THE AI TOOLCHAIN

NO. 028

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED SEPTEMBER 16, 2026 · EVERY DAY

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

▸ // HOW THIS ISSUE IS MADE SHOW

VIEW

ISSUE VIEW

scroll



SCROLL | EXPAND IN PLACE | PREVIEW PANE

INFORMATION DENSITY

everything



BRIEFED | EVERYTHING

Every tool in the issue, in full.

all 245 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 47 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. ai-gateway now drops PII, secrets and prompt-injection filters into the call path as gateway plugins, so no as to implement them. Pinecone replaces its records model with a

vim ? help

schema-based Documents API. And Cursor's Projects hands long-running work to a coordinator that fans it out across parallel cloud subagents.

GOVERN

Strip secrets and PII out of prompts without touching application code

ai-gateway v1.5.7's `pii-redact`, `secret-scan`, `schema-guard`, `regex-guard` and `prompt-shield` plugins run on the hop every model call already makes, so filtering stops being a per-service library each team re-implements and drifts on. FailproofAI 1.0.6 extends the same idea to tool calls, enforcing policy before a Hermes agent acts rather than auditing after it did.

ai-gateway · FailproofAI

BUILD

Model a RAG corpus as typed documents instead of flat records

Pinecone's Documents API adds schema-based indexes, prefix-based listing and partial field patching, so updating one field on a chunk no longer means re-upserting the whole record. Milvus's C++ SDK finally gets multi-vector hybrid search with full-text sparse sub-requests and RRF reranking, closing the gap for C++ services that had to shell out to another language to combine dense and keyword retrieval.

Pinecone · Milvus

BUILD

Hand off a multi-day refactor and let it run in parallel while you sleep

Cursor Projects puts a coordinator agent in front of thousands of parallel cloud and local subagents, triggered from Slack, PRs, CI or a schedule, with `/goal` carrying a long-lived objective across runs — work that previously had to be babysat in one interactive session. Mastra 1.67 adds per-call background execution dispositions with completion signals and transactional thread ownership transfer, so a handoff between agents cannot leave a thread half-owned.

Cursor · Mastra

OFFENSIVE SECURITY

◆ Exploitation & C2

xalgorix

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

The xalgorix platform runs AI pentesting agents for reconnaissance, vulnerability detection, and exploitation workflows.

> xalgorix v4.6.77 focuses on reliability fixes for its LLM-driven agent pipeline: preserving multi-part model identifiers, fixing subagent configuration inheritance, tightening proof-integrity checks during vulnerability verification, and correcting proxy header handling.

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

01 **Hop-by-hop header stripping for proxies** IMPROVED 60

Strips hop-by-hop HTTP headers `Proxy-Connection`, `Connection`, and `Keep-Alive` to maintain proxy protocol integrity.

– Names exact headers stripped, a concrete technical fix v4.6.77

THINNER COVERAGE BELOW

02 **Slash-separated model ID preservation** IMPROVED 55

Preserves slash-separated model identifiers such as `zai-org/GLM-5.3`, `meta-llama/Llama-3.1-70B-Instruct`, and `deepseek-ai/DeepSeek-V3` when routing through custom OpenAI-compatible endpoints or cloud gateways.

– Names exact model IDs but no config surface or command shown v4.6.77

03 **Subagent LLM configuration propagation** IMPROVED 50

Propagates cloned LLM client configuration, temperature settings, and custom resolvers directly to delegated subagents, preventing 404 Model Not Found errors on child tasks.

– Explains mechanism and fixed error but no config keys named v4.6.77

04 **Proof-integrity enforcement in vulnerability verification**

NEW 35

Enforces proof-integrity rules forbidding LLMs from synthesizing or hallucinating HTTP response bodies during vulnerability verification.

– States the rule but no mechanism or configuration detail v4.6.77

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

StackBlitz bolt.new

Feedback ⓘ 1 RELEASE · SEEN 2026-09-16

NOTES ↗ RANK

Bolt.new is an AI-powered web development platform that builds websites and applications from prompts.

> Bolt.new introduced a third AI agent, Bolt Forge, running open-source models with a larger uncapped-daily allowance, alongside a new $9/month Bolt Lite plan, in-preview visual editing via a Select tool, and a round of Bolt Slides authoring and export additions.

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

01 **Visual in-preview editing with Select tool** NEW 90

A new Select tool in the bolt.new chatbox lets you click an element in the preview to edit its text, color, or font size directly, make multiple such edits, then click Save changes to write all edits to project code in one token-consuming step rather than per-edit.

Iterate on UI styling without consuming tokens — use the Select tool to batch visual edits, then commit them all at once.

> 📍 In the bolt.new chatbox, click the Select tool, click an element in the preview to edit its text, color, or font size, make additional edits as needed, then click Save changes to write all edits to project code in one token-consuming step.

— Exact tool name, UI steps, and token-saving mechanism described. snapshot-20260916

02 **Bolt Slides authoring and export tools** NEW 70

Bolt Slides gained PDF and JSON export for sharing decks as files or importing into other projects; direct slide editing to change text, colors, and formatting without prompting; slide organization controls to reorder, duplicate, or delete slides from the side panel or grid overview; and a speaker view showing notes, the next slide, and a timer in the presenter's tab while the audience sees a fullscreen view.

— Names every addition though no navigation path given. snapshot-20260916

THINNER COVERAGE BELOW

03 **Bolt Lite standalone plan** NEW 50

Bolt Lite is a new $9/month standalone plan that provides Forge-only access to the core Bolt experience.

— Names price and scope but no signup steps given. snapshot-20260916

WAS THIS USEFUL? 👍 👎

Anysphere Cursor

Feedback ⓘ 1 RELEASE · SEEN 2026-09-16

NOTES ↗ RANK

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

> Cursor's biggest addition this window is Projects: a coordinator agent that plans and delegates long-running work to thousands of parallel cloud and local subagents, running autonomously via Slack, PR, CI, or scheduled triggers, and now runs on Cloud Agents by default. It also ships per-Project shared context, a `/goal` command for long-lived objectives, and steering tweaks for follow-up messages.

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

01 **`/goal` command for long-lived objectives** NEW 80

The `/goal` command gives an agent a long-lived objective to pursue until fully complete, for example `/goal fix all flaky tests and make CI green`.

Kick off a long-running CI health objective so the agent keeps working on flaky tests across sessions without needing to be re-prompted.

📍 `/goal` fix all flaky tests and make CI green

— Exact command shown with a runnable example snapshot-20260916

02 **Subscriptions for autonomous triggers** NEW 75

Inside a Project, the coordinator can be told to watch a Slack channel, follow all PRs, track CI runs, or run on a recurring schedule, delegating work automatically based on detected signals instead of waiting for a user prompt. Set up via the Agents Window > Projects > New Project, then instruct the coordinator in chat (e.g. 'Watch the #bug-reports Slack channel and delegate a fix to an agent each time a new bug is posted'); the resulting subscription can be reviewed or removed via the Listening pill above the chat input.

Set up a Project to automatically triage and delegate fixes from a Slack bug-report channel without manual prompting.

📍 1. In the Agents Window, find the Projects section in the sidebar and click New Project. 2. Name the project, choose the target GitHub repository under Workspace, and select a coordinator model. 3. Click Create Project and, in the chat with the coordinator, type: 'Watch the #bug-reports Slack channel and delegate a fix to an agent each time a new bug is posted.' 4. The coordinator creates the

subscription. Click the Listening pill above the chat input to review or remove the subscription.

Connect a Slack bug-report channel to a Project so the coordinator automatically delegates a fix each time a new bug is reported, with no manual prompting.

📍 In Cursor, open Projects from the left-hand nav › select your Project › open Subscriptions › connect Slack › point it at your bug-report channel.

— Exact triggers and step-by-step setup given `snapshot-20260916`

---

03 **Projects: coordinator agent for long-running work** NEW

65

Projects is a new long-running work unit in the Agents Window (beta), accessible from the left-hand nav, where a coordinator agent plans work and delegates to parallel cloud and local subagents — thousands of them for large bodies of work like features, migrations, or full apps — maintaining shared context across all of them over months, without requiring a prompt at every step.

Set up a Project to automatically triage and delegate fixes from a Slack bug-report channel without manual prompting.

📍 1. In the Agents Window, find the Projects section in the sidebar and click New Project. 2. Name the project, choose the target GitHub repository under Workspace, and select a coordinator model. 3. Click Create Project and, in the chat with the coordinator, type: 'Watch the #bug-reports Slack channel and delegate a fix to an agent each time a new bug is posted.' 4. The coordinator creates the subscription. Click the Listening pill above the chat input to review or remove the subscription.

— Names UI location and mechanism, but no CLI/API surface `snapshot-20260916`

THINNER COVERAGE BELOW

---

04 **Cloud Agents run Projects by default** NEW

55

Projects now run on Cloud Agents by default so work continues after the local machine is closed, with the coordinator dispatching a local agent only when something needs testing on the user's machine. Related screenshots show a Remote Machines menu with self-hosted options (lambda-test, cloud-demo, jacks-desktop) and a Cloud Agents dashboard listing Self-hosted Machines pools with active and idle worker counts.

— Behavior described with supporting screenshots, no exact command `snapshot-20260916`

---

05 **Steering behavior for follow-up messages** IMPROVED 50

Follow-up messages sent while an agent is working now wait for the next tool call instead of interrupting mid-action; users can send immediately by clicking Send now or pressing Enter twice.

— Clear before/after behavior with a minor UI action snapshot-20260916

06 **Shared context across Project agents** NEW 45

Each Project has a shared context: a set of files that sync across every cloud and local agent in it, so knowledge one agent gains — test instructions, codebase conventions, preferences — is available to every future agent working in that Project.

— Mechanism described but no usage steps given snapshot-20260916

WAS THIS USEFUL? 👍 👎

GitHub Copilot CLI

Feedback ⓘ 1 RELEASE · 2026-09-16

NOTES ↗ RANK

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

> Copilot CLI v1.0.85 restructures the plugin/mcp/skill/lsp subcommands and changes `COPILOT_ALLOW_ALL` semantics as breaking changes, while adding Vim modal editing, a `/config` sidebar, `/sandbox` network allow/deny rules, GPT-6 Astra model support, and JSON output flags for plugin commands.

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

01 Plugin/MCP/skill CLI subcommand restructuring

BREAKING

85

`copilot plugins list --kind instruction` and `--kind lsp` are replaced by new `copilot instruction list` and `copilot lsp list` subcommands, and `copilot plugins enable/disable --plugin|--mcp|--skill` is replaced by dedicated `enable` and `disable` subcommands on `copilot plugin`, `copilot mcp`, and `copilot skill`.

Enable a specific MCP server without dropping into interactive mode — useful in onboarding scripts.

```
$ copilot mcp enable my-server
```

— Names old and new commands precisely with example. v1.0.85

02 JSON output for plugin commands NEW

80

Adds `--json` flag to `copilot plugin list`, `copilot plugin marketplace list`, and `copilot plugin marketplace browse` for machine-readable output, enabling scripted pipeline checks or dashboards.

Get structured JSON output from the plugin marketplace for scripted pipeline checks or dashboards.

```
$ copilot plugin marketplace list --json
```

— Exact flag and commands with runnable example. v1.0.85

03 Vim modal editing mode NEW

80

Adds `editorMode` set to `vim` (or the `/vim` slash command) to enable modal editing in the composer, with the current mode shown while typing.

Switch to Vim modal editing in every session by persisting the preference in config.

```yaml
editorMode: vim
```

— Exact config key and slash command with example. v1.0.85

---

04 **COPILOT_ALLOW_ALL falsey value behavior change** BREAKING

75

`COPILOT_ALLOW_ALL` now accepts values such as `1`, `0`, `yes`, or an empty string, and falsey values now disable automatic tool approval instead of enabling it, breaking setups that relied on a falsey value meaning `allow all`.

— Exact env var and before/after behavior described. v1.0.85

---

05 **allowManagedHooksOnly policy for extension hooks** NEW

75

Adds `allowManagedHooksOnly` policy coverage to block extension-registered `preToolUse`, `postToolUse`, and `postToolUseFailure` callbacks.

— Exact policy key and three named callback types. v1.0.85

---

06 **Worktree commands out of experimental mode** IMPROVED

65

Makes `/worktree`, `/move`, and `--worktree` available without enabling experimental mode.

— Exact commands with clear before/after change. v1.0.85

---

07 **Concise transcript view** IMPROVED

60

Sets `transcriptView` to `'concise'` to group tool activity into expandable work summaries.

— Exact config key and value with brief mechanism. v1.0.85

THINNER COVERAGE BELOW

---

08 **/sandbox network allow/deny rules** NEW

55

Adds `/sandbox` network host allow/deny rules that can be layered without replacing your configured upstream proxy.

— Names command and behavior, no config keys given. v1.0.85

09 **Agent Factory run pause/resume controls** NEW 55

Adds pause and resume controls for Agent Factory runs from the `/factories` dialog.

— Names dialog and controls, no further mechanism. v1.0.85

10 **/config sidebar configuration screen** NEW 50

Adds the `/config` slash command to open a sidebar configuration screen directly in the CLI.

— Names slash command but gives no further options. v1.0.85

11 **/settings additions for taskbar and context tools** NEW 50

Adds `/settings` options including `taskbarPresence false` to disable Windows taskbar session status, and an opt-in to context management tools for agents and subagents.

— Names one exact flag; second option lacks a name. v1.0.85

12 **Automatic compaction checkpoints** NEW 50

Adds an automatic compaction checkpoint that appears in `/session checkpoints`.

— Names surface but limited mechanism detail. v1.0.85

13 **Indexed search reliability improvements** IMPROVED 50

Indexed search now works on Windows ReFS volumes, supports explicit cloud-sync overrides, and keeps refreshing on Linux when native file watches are exhausted.

— Names platforms and conditions, no config keys. v1.0.85

14 **Session and memory import for JSONL format** NEW 40

Adds session and memory import commands for the semantic JSONL interchange format.

— No exact command name or flag given. v1.0.85

15 **Windows artifacts drop VC++ Redistributable requirement** IMPROVED 40

Windows CLI artifacts now run without requiring the Visual C++ Redistributable.

— Simple environmental fix, no further detail. v1.0.85

16 **GPT-6 Astra model support** NEW 35

Adds support for the GPT-6 Astra model.

— Bare mention, no usage detail. v1.0.85

17 **Plugin-contributed agent selection** NEW 35

Enables plugin-contributed agents discovered by the CLI to be selected and run.

— No command name or mechanism given. v1.0.85

18 **/copy includes task completion messages** IMPROVED 35

Enables `/copy` to include task completion messages when available.

— Thin, single-line addition. v1.0.85

19 **Scheduled prompts shown in footer** IMPROVED 25

Shows active scheduled prompts in the CLI footer by default.

— Bare behavior change, no configuration detail. v1.0.85

└──▷ **BREAKING ON UPGRADE**

! `copilot plugins list --kind instruction` and `--kind lsp` are replaced by `copilot instruction list` and `copilot lsp list`.

! `copilot plugins enable/disable --plugin|--mcp|--skill` is replaced by `enable` and `disable` subcommands on `copilot plugin`, `copilot mcp`, and `copilot skill`.

! `COPILOT_ALLOW_ALL` falsey values (such as `0`, `no`, or empty string) now disable automatic tool approval instead of enabling it; setups relying on a falsey value to mean `allow all` will break.

WAS THIS USEFUL? 👍 👎

Letta Code

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

Letta Code is a terminal coding agent with persistent memory and identity across development sessions.

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

01 **Subagent naming as shadows of parent** IMPROVED 30

Subagents are now named as shadows of their parent agent, making multi-agent task hierarchies easier to trace and identify.

— Explains the change but no naming scheme or command shown. v0.32.11

WAS THIS USEFUL? 👍 👎

OpenAI Codex CLI

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗

CODE ↗ RANK

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

> Codex CLI added a way to pin the daemon to the exact CLI package in use, made forked conversation threads inherit their attachments, and brought display math rendering to the TUI.

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

01 **Daemon package replacement via --from-cli** NEW 90

Adds `codex app-server daemon update --from-cli` to copy and pin the invoking CLI's complete package as the daemon, including downgrades and local builds. Proceeding requires interactive confirmation or the `--yes` / `-y` flag to confirm non-interactively (which itself requires `--from-cli`).

Replace a running daemon with the exact CLI package you have installed locally — useful after a manual install or downgrade — without being prompted interactively.

```
$ codex app-server daemon update --from-cli --yes
```

— Exact flags and a runnable command are given rust-v0.155.0-alpha.7

02 **Thread attachment inheritance for forks** NEW 80

Copies current thread attachments into non-ephemeral conversation forks. Clients can retrieve the inherited attachments via `thread/attachment/list` using `forkedFromId` from the `thread/started` event.

— Endpoint and field named but no usage example rust-v0.155.0-alpha.7

THINNER COVERAGE BELOW

03 **Display math rendering in the TUI** NEW 10

The TUI now supports display math rendering.

— Only named in the summary, no mechanism given rust-v0.155.0-alpha.7

WAS THIS USEFUL? 👍 👎

Diagram Design

Feedback ⓘ changes since 2026-08-17 CODE ↗

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

Diagram Design: Adds Excalidraw diagram import for editorial redraw.

└──▷ **GET THIS VERSION**

```
$ git clone --branch commits-2026-08-17
https://github.com/cathrynlavery/diagram-design.git
# already have the repo? check out this version:
$ git checkout commits-2026-08-17
```

└──▷ **TRY IT**

Validate a sankey diagram's geometric contracts — confirming the checker fires on defects and stays quiet on clean files — using the new adversarial test suite.

```
$ python3 scripts/test-verify-sankey.py
```

Adds native Droid plugin packaging via the factory.

Adds `lint-render.py`, a headless Chromium render-based linter that catches clipped SVG content, collapsed SVGs, sideways page overflow, and runtime errors unreachable by source-only linting.

Adds adversarial tests for `verify-sankey.py` covering both polarities — firing and staying quiet.

WAS THIS USEFUL? 👍 👎

Anthropic Claude Code

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ CODE ↗ RANK

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

> Claude Code shipped a standalone app with visual diff review and cloud sessions, alongside a v2.1.273 release adding LLM gateway hint headers, MCP disconnect alerts, remote-control session forking, richer OTEL metrics, and a switch to local auto-mode classification.

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

01 **LLM gateway hint headers** NEW 90

Adds `x-claude-code-request-class`, `x-claude-code-agent-type`, `x-claude-code-prev-tool-durations`, `x-claude-code-compaction`, and `x-claude-code-context-compacted` request headers for LLM gateways; opt in with `CLAUDE_CODE_GATEWAY_HINT_HEADERS=1`.

Send routing context to a corporate LLM gateway so it can classify, prioritize, or log Claude Code requests by agent type and compaction state.

```
$ CLAUDE_CODE_GATEWAY_HINT_HEADERS=1 claude
```

— Names exact headers and env var to enable. v2.1.273

02 **Remote-control session forking** NEW 75

Adds the ability to fork a session started with `claude --remote-control` or `/remote-control` from the Claude app, with the fork running as a background session on your local machine.

— Names exact commands and resulting behavior. v2.1.273

03 **Local auto-mode classifier default** BREAKING 75

The default auto-mode classification switches from server-side to a local classifier; restore server-side classification on Bedrock, Vertex, or Foundry with `CLAUDE_CODE_AUTO_MODE_SERVER=1 claude`.

Restore server-side auto-mode classification on Bedrock, Vertex, or Foundry after the default switches to the local classifier in this release.

```
$ CLAUDE_CODE_AUTO_MODE_SERVER=1 claude
```

— Gives exact env var to revert the new default. v2.1.273

04 **OTEL tool detail metrics naming** IMPROVED 65

Changes `OTEL_LOG_TOOL_DETAILS=1` to also include real agent, skill, plugin, and MCP server names on cost and token metrics.

— Names exact env var and metric scope affected. v2.1.273

THINNER COVERAGE BELOW

05 **MCP server disconnect notification** NEW 50

Adds a notification when an MCP server disconnects mid-session and automatic reconnection gives up, pointing at `/mcp` for remediation.

— Names the /mcp remediation path but little else. v2.1.273

06 **Standalone app with visual diff and cloud sessions** NEW 45

Claude Code now ships a standalone app for running outside an IDE or terminal, offering visual diff review, multiple side-by-side sessions, scheduled recurring tasks, and cloud session launch.

— Lists new app capabilities but no launch command or config. product docs

WAS THIS USEFUL? 👍 👎

Graphify Labs Graphify

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

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

> Graphify's latest releases add Terraform local-module topology (with a required post-upgrade regeneration step), Ruby inheritance-chain resolution, and precise Markdown-to-symbol reference edges.

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

01 **Terraform local-module topology and required graph regeneration**

80

BREAKING

Graphify now builds topology for Terraform configurations that use local module sources, including reworked Terraform node ids. After upgrading an existing graph, users must run the regeneration command once to rebuild Terraform ids and topology (tracked in issue #3571); this is done by running `graphify update .` against the repo.

After upgrading, regenerate Terraform node ids and caller→implementation topology for a repo that uses local module sources.

```
$ graphify update .
```

— Names the exact regeneration command and the reason it's required. v0.9.62

THINNER COVERAGE BELOW

02 **Markdown inline-code reference edges** NEW

55

Graphify now emits a `references` edge from an inline code span in a Markdown file (e.g. `Widget.render`) to the symbol it names, but only when exactly one corpus callable matches the name, which prevents false edges from ambiguous references.

— Mechanism and disambiguation rule described, but no user-facing command. v0.9.62

03 **Ruby inheritance-chain resolution** NEW

15

Graphify adds resolution of Ruby inheritance chains when building the knowledge graph.

— Only named, no mechanism or scope given. v0.9.62

└──▷ BREAKING ON UPGRADE

! After upgrading an existing graph, you must run the regeneration command once to rebuild Terraform ids and topology (see #3571).

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗

RANK

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

> Qwen Code's v0.24.0 release expands its workflow engine with named, budget-aware workflow runs, adds a multimodal media pipeline and DingTalk shared output, and rolls out Goal lifecycle tracking, hooks tooling, and web shell UI upgrades — alongside a breaking change to variable expansion in command hooks.

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

01 Workflow engine: named runs, budgets, and grants NEW

85

Extensions can now ship dynamic workflows that the model invokes when they declare `whenToUse`. Workflows support a `+Nk` directive to share a per-turn token budget across steps, and the `serve` subsystem can `start workflow runs` with `args`, `sourceRef`, and a caller-supplied script. Saved workflows can be run by name, with grants pinned to script content; the engine also correlates agent dispatches and nested workflow calls, preserves source references across workflow runs and resumes, and adds workflow size guidelines, a structure preview, and large-run warnings.

— Names directives, subcommand and fields; runnable but no full example. v0.24.0

02 Hooks dialog and progress events NEW

65

Adds a full `/hooks` dialog in the OpenTUI CLI interface, with live registry reload when `/hooks` opens, and emits hook progress events from the core runtime.

— Names UI command and behaviour; clear starting point. v0.24.0

03 Goal lifecycle tracking and budgets NEW

65

Emits a Goal lifecycle event and associated metrics via telemetry, bills foreground subagents and verifier token usage against the Goal, shows turn and active-time budgets in the Goal view, and tracks background result execution across daemon and web shell.

— Describes mechanism and surfaces across daemon and web shell. v0.24.0

04 Web shell UI enhancements NEW

60

Adds footnote previews and per-turn sources in the web shell, file preview from expanded tool details, the ability to connect the web shell to a selected remote daemon, and customizable context details with collapsible snapshots.

— Several named UI additions, no exact navigation path. v0.24.0

THINNER COVERAGE BELOW

05 **DingTalk shared output via channels subsystem** NEW

55

Adds shared output modes with DingTalk support via the channels subsystem, including sender identity in direct messages.

— Names integration and subsystem but no config path. v0.24.0

06 **Breaking: bash expands directory variables in hooks** BREAKING

55

Bash now expands project directory variables in command hooks — hooks that relied on literal dollar-sign strings in those positions will behave differently after upgrade.

— Clear before/after but no exact variable names listed. v0.24.0

07 **Budget-based ACP child admission in serve subsystem** NEW

45

Adds budget-based ACP child admission and idle-child reclamation to the `serve` subsystem.

— Mechanism named but no exact command or flag. v0.24.0

08 **Per-session cap on web_search calls** NEW

40

Caps `web_search` tool calls per session to limit runaway search usage.

— No limit value or config key given. v0.24.0

09 **Summary projection for prompt and replay APIs** NEW

35

Supports summary projection for prompt and replay APIs in the daemon.

— No endpoint or field names given. v0.24.0

10 **Multimodal media pipeline (omni)** NEW

35

Integrates a multimodal media pipeline (`omni`) into the main branch.

— Named but undescribed beyond integration. v0.24.0

11 **External reasoning profiles in core** NEW

25

Supports external reasoning profiles and defaults in the core.

— Bare description with no mechanism or config. v0.24.0

12 **Local enumeration of installed extension skills** NEW

25

Enumerates installed extension skills locally from the daemon.

— Bare capability with no interface detail. v0.24.0

└──▷ **BREAKING ON UPGRADE**

! Bash now expands project directory variables in command hooks — hooks that relied on literal dollar-sign strings in those positions will behave differently after upgrade.

WAS THIS USEFUL? 👍 👎

Charm Crush

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗ CODE ↗

RANK

Terminal-based AI coding assistant that integrates LLMs, LSPs, and MCP servers into your development workflow.

> Crush v0.95.0 introduces Plan Mode, a read-only agent that explores a codebase and proposes changes without writing files.

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

01 **Plan Mode read-only agent** NEW 60

A new read-only agent mode that reads files, greps, maps architecture, checks tests and config, and spawns sub-agents to explore a codebase — without writing a single file or running a command. Users can review, refine, or approve the plan before any code changes are made.

— Explains mechanism and workflow but no command or flag to invoke it v0.95.0

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

Langflow

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗ RANK

Langflow is a visual builder for agents and RAG workflows that exports each flow as an API or Python object for deployment.

> Langflow v1.12.2 upgraded the Guardrails component to support combined rule-based and model-based validation checks in a single guard.

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

01 **Combined checks in Guardrails component** IMPROVED 40

The Guardrails component now supports optional combined rule and model checks, allowing both rule-based and model-based validation to run together within a single guard.

— Names the mechanism but no config keys or examples given. v1.12.2

WAS THIS USEFUL? 👍 👎

AWS Strands Agents

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

Strands Agents is AWS's open-source SDK for production agents in Python and TypeScript, with a model-driven loop, tool integration, and deployment targets.

> Strands Agents' TypeScript and Python SDKs picked up new vended tools (web_fetch, notebook), retrieval upgrades (BM25 search, context strategy presets), Python-side background task and session-manager support, deeper bidirectional streaming (live transcripts, completion events, shared hooks), and a non-clobbering seam for Anthropic-direct server-side tools.

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

01 **backgroundTasks on the Python Agent class** NEW 65

Adds `backgroundTasks` to the Python `Agent` class for running tasks outside the main agent loop.

– Names the exact attribute and its purpose, a clear starting point. typescript/v1.18.0

THINNER COVERAGE BELOW

02 **Shell tool result improvements: exit codes and partial output** IMPROVED 59

Shell tool results now include `exit_code` so agents and callers can inspect the process exit status directly, and shell timeout errors now include partial output so callers can see what was produced before the timeout occurred.

– Names the exact field but no example invocation. typescript/v1.18.0

03 **Non-clobbering seam for Anthropic-direct server-side tools** NEW 55

Adds a non-clobbering seam for Anthropic-direct server-side tools (such as web search), preserving custom tool configs when Anthropic-managed tools are also active.

– Explains the mechanism and effect but no exact API named. python/v1.56.0

04 **New vended tools: web_fetch and notebook** NEW 53

Adds a `web_fetch` tool for TypeScript agents, enabling HTTP fetch as a vended tool alongside the existing shell and notebook tools, and ports the notebook tool to both Python and TypeScript as a vended tool.

– Names the tools but no usage example or config shown. typescript/v1.18.0

05 **Context strategy presets and BM25 search** NEW 50

Adds context strategy presets for TypeScript with rewired defaults to use class-based strategies, and adds a BM25 search strategy to the storage layer as an alternative retrieval method.

— Describes the additions but no config keys or example given. typescript/v1.18.0

06 **Bidirectional streaming improvements** IMPROVED 50

Bidirectional (bidi) streaming now renders live transcripts with audio output and adds transcript completion events, shares lifecycle hooks across bidi streaming with new response completion hooks, and aligns model configuration and parameter overrides across the bidi path.

— Several concrete behavior changes but no named API or config. typescript/v1.18.0

07 **OpenAI prompt-cache keys from session ID** NEW 38

Automatically derives OpenAI prompt-cache keys from the session ID.

— Names the mechanism but no config or code path given. typescript/v1.18.0

08 **Session manager integration for Python context manager** NEW 21

Adds session manager integration for the Python context manager.

— Single-line description with no mechanism or surface named. typescript/v1.18.0

WAS THIS USEFUL?  

LiveKit Agents

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗

RANK

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

> LiveKit Agents added native support for nine realtime speech-to-speech model providers plus a half-cascade architecture option, introduced EU data-residency controls for LiveKit Inference, and shipped a batch of plugin, telemetry and performance improvements including a Synthesia avatar plugin and expanded Agent Insights tracing.

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

01 Realtime model plugin support and half-cascade architecture

NEW 95

Adds `RealtimeModel` plugin support for Amazon Nova Sonic, Azure OpenAI Realtime API, Gemini Live API, NVIDIA PersonaPlex, OpenAI GPT-Live, OpenAI Realtime API, Phonic Speech-to-speech, SpaceXAI Grok Voice Agent API, and Ultravox Realtime, usable in place of an LLM instance inside `AgentSession` or `Agent` by passing the model to the `llm` constructor parameter (Python and Node.js), bypassing the need for an STT-LLM-TTS pipeline. Full-duplex models such as OpenAI GPT-Live simultaneously listen and speak, with turn boundaries calculated from audio by the framework rather than the caller. Also supports `modalities=['text']` on realtime model plugins to enable a half-cascade architecture, pairing realtime speech comprehension with a separate `tts` instance (e.g. `'inworld/inworld-tts-2'`) passed to `AgentSession`.

Use a realtime model with a separate TTS for full control over speech output while still benefiting from realtime audio comprehension (half-cascade architecture).

python

```python
from livekit.agents import AgentSession
from livekit.plugins import openai

session = AgentSession(
    llm=openai.realtime.RealtimeModel(modalities=["text"]),
    tts="inworld/inworld-tts-2"
)
```

Drop a realtime model directly into an AgentSession to skip the STT-LLM-TTS pipeline entirely and process speech end-to-end.

```python
from livekit.agents import AgentSession
from livekit.plugins import openai

session = AgentSession(
    llm=openai.realtime.RealtimeModel()
)
```

— Names every provider and the exact constructor param, with runnable code. product docs

02 **EU data residency controls for LiveKit Inference** NEW

83

Adds an Inference region restriction setting under Project → Settings → Data and privacy in LiveKit Cloud to restrict inference routing to the project's data region (e.g., EU only), with EU-hosted projects defaulting to enabled for new projects. Also exposes an EU endpoint column in the LLM, speech-to-text, and text-to-speech model tables in Agent Builder, showing which models have a dedicated EU endpoint via LiveKit Inference.

Restrict a pre-existing EU-region LiveKit project's inference traffic to EU-only model endpoints to satisfy data residency requirements.

📍 In LiveKit Cloud, go to Project → Settings → Data and privacy and set Inference region restriction to Enabled.

— Exact settings path given, but only UI steps not a command. product docs

THINNER COVERAGE BELOW

03 **Agent Insights telemetry expansion** IMPROVED 55

Traces RPC calls made and handled by agents, surfaced in Agent Insights telemetry. Lifecycle tracing is extended to cover dispatch, startup, shutdown, end-of-turn wait, and turn handoff. Also detects asyncio event-loop blocking and surfaces the offending call in Agent Insights.

— Lists specific traced lifecycle stages but no query or API path. livekit-agents@1.8.2

04 **STT plugin updates: ElevenLabs and Speechmatics** IMPROVED

44

ElevenLabs realtime STT gains secondary-language support and automatic language detection. The Speechmatics STT plugin is migrated to the Agent STT interface.

— Names the two plugins but no config or usage detail. livekit-agents@1.8.2

05 **Prefork startup preloading** IMPROVED 40

Preloads libraries in the prefork server before forking, cutting agent startup time by up to 800ms.

— Concrete number given but no configuration or opt-in step. `livekit-agents@1.8.2`

06 **room_id in IPC job logging** IMPROVED 40

Adds `room_id` to IPC job logging extras for richer per-job log context.

— Names the exact logging field but nothing else. `livekit-agents@1.8.2`

07 **Synthesia interactive-avatar plugin** NEW 33

Adds a Synthesia interactive-avatar plugin for avatar-driven agent sessions.

— One-line mention with no mechanism or setup steps. `livekit-agents@1.8.2`

08 **New Gemini Live models** NEW 23

Adds new Gemini Live models to the Gemini Live backend.

— Bare mention, no model names or migration notes. `livekit-agents@1.8.2`

WAS THIS USEFUL? 👍 👎

nanobot

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

Nanobot is a self-hosted personal AI assistant framework with tools, memory, scheduled tasks, and integrations with chat applications.

> nanobot v0.3.5 ships a native terminal UI and a new browser webui workbench, reworks gateway lifecycle and session storage with several breaking migrations, and adds new model/gateway integrations, sandboxing options, and chat-channel improvements.

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

01 Concurrency and stream timeout controls

BREAKING 95

Adds `NANOBOT_MAX_CONCURRENT_REQUESTS` environment variable to cap concurrent gateway requests (unlimited unless set to a positive integer), and `NANOBOT_STREAM_IDLE_TIMEOUT_S` (default 90 seconds) for a per-event-renewed model stream idle timeout, replacing the old fixed total-duration timeout. The old `agents.defaults.failOnToolError` config field and `NANOBOT_LLM_TIMEOUT_S` environment variable are no longer used.

Cap concurrent agent requests and tune stream idle timeout to protect a shared gateway under load.

```
$ NANOBOT_MAX_CONCURRENT_REQUESTS=20
NANOBOT_STREAM_IDLE_TIMEOUT_S=120 nanobot gateway --background
```

— Env vars, default value, and deprecations named with example v0.3.5

02 Session storage migration and rollback

BREAKING 90

Migrates session JSONL storage to `sessions/<workspace-id>/` under the configuration directory on first upgraded start; old and new versions must not write the same workspace concurrently. Before downgrading, run `nanobot sessions restore-workspace --config <config.json> --workspace <workspace>` to roll back session storage — the command refuses to overwrite a different existing file.

— Exact migration path and rollback command given v0.3.5

03 Email authentication trust anchors

BREAKING 90

Adds `channels.email.trustedAuthservIds` config key to enforce explicit receiver trust anchors when `verifySpf` or `verifyDkim` is enabled. Breaking: existing verified

Email configurations with `verifySpf` or `verifyDkim` enabled but lacking `trustedAuthservIds` will fail to start.

Enforce strict email authentication in a multi-tenant deployment by naming the trusted authserv identity your mail relay stamps onto every inbound message.

```yaml
channels:
  email:
    verifySpf: true
    verifyDkim: true
    trustedAuthservIds:
      - mail.example.com
```

— Config key, example, and failure condition all specified v0.3.5

---

04 **Native terminal UI for nanobot** BREAKING 85

The `nanobot` command now opens a native terminal UI (replacing the old Python prompt) with session switching, clipboard image paste, app/skill tab-completion, context and file-diff inspection, follow-up queuing, and `/detach` to keep the gateway running headless after closing the client. Use `nanobot --classic` for the legacy prompt or `nanobot -m "<message>"` for one-shot invocations.

— Names flags and default-behavior change but no command example v0.3.5

---

05 **Gateway lifecycle control** BREAKING 85

Adds `nanobot gateway --background` to keep the gateway alive after all interactive clients disconnect. As a breaking change, the last attached interactive WebUI/TUI launcher now owns gateway lifetime — closing it stops the gateway, so `/detach` or `nanobot gateway --background` must be used to keep it running.

Keep the gateway running in the background so agent tasks continue after you close your terminal session.

```
$ nanobot gateway --background
```

— Runnable command plus clear before/after behavior v0.3.5

06 **Gateway health readiness reporting** BREAKING 70

`/health` now returns HTTP 503 when its enabled WebSocket channel is not running; monitoring that assumed every live process returned 200 must be updated.

— Named endpoint with exact status code behavior v0.3.5

THINNER COVERAGE BELOW

07 **webui subcommand for browser workbench** NEW 55

Adds `nanobot webui` subcommand to launch a browser-based workbench that shares the same gateway and saved WebSocket conversations as the terminal client.

— Command named but no further mechanism given v0.3.5

08 **New model and gateway integrations** NEW 50

Adds support for DeepSeek Responses and V4 vision models, Eden AI and OrcaRouter gateways, and AnySearch.

— Names integrations but no configuration detail v0.3.5

09 **Subprocess sandboxing improvements** IMPROVED 50

Adds optional macOS Seatbelt isolation for subprocess sandboxing and allows extra `bwrap` bind roots in sandbox execution.

— Names sandbox mechanisms but no config keys given v0.3.5

10 **SkillHub, Agent Plugins, and OAuth model catalog** NEW 45

Adds SkillHub discovery, portable Agent Plugins, and an online OAuth model catalog.

— Three additions named but none explained further v0.3.5

11 **Automations task and calendar views** NEW 45

Adds task and calendar views for Automations, with per-run reply inspection and clearer scheduling controls.

— Describes UI additions without exact navigation v0.3.5

12 **Chat channel improvements across Telegram, Feishu, DingTalk** IMPROVED 45

Adds rich Telegram streaming, improved Feishu onboarding, and DingTalk private-chat gating with mention-sender behaviour in group replies.

— Names three channels but no config specifics v0.3.5

13 **Remote MCP OAuth and token refresh** NEW 40

Adds browser OAuth and token refresh for remote MCP servers.

— Describes capability but not how it's configured v0.3.5

14 **Context usage visibility** IMPROVED 35

Adds per-round token/cache context charts and context-compaction progress visibility.

— Feature named without location or config v0.3.5

15 **SDK host integration extension points** NEW 25

Adds SDK host integration extension points for embedding nanobot in larger applications.

— Bare mention with no mechanism or API named v0.3.5

16 **Configurable idle compaction scan interval** IMPROVED 25

Adds a configurable idle compaction scan interval for the agent.

— No config key or default value provided v0.3.5

17 **PWA home-screen installation** NEW 25

Adds PWA home-screen installation support for mobile users.

— One-line mention with no setup detail v0.3.5

└──▷ **BREAKING ON UPGRADE**

! Session JSONL files now live under `sessions/<workspace-id>/` inside the configuration directory; the first upgraded start migrates existing workspace sessions — old and new versions must not write the same workspace concurrently.

! Existing verified Email configurations that have `verifySpf` or `verifyDkim` enabled but lack `channels.email.trustedAuthservIds` will not start.

! Plain `nanobot` now opens the native terminal UI instead of the Python prompt; use `nanobot --classic` for the old prompt or `nanobot -m` for one-shot use.

! The last attached interactive WebUI/TUI launcher now owns gateway lifetime — closing it stops the gateway; use `/detach` or `nanobot gateway --background` to keep it running.

! The old `agents.defaults.failOnToolError` config field and `NANOBOT_LLM_TIMEOUT_S` environment variable are no longer used.

! `/health` can now return HTTP 503 when the enabled WebSocket channel is not running; monitoring that assumed every live process returned 200 must be updated.

! Before downgrading, stop nanobot and run `nanobot sessions restore-workspace --config <config.json> --workspace <workspace>`; the command refuses to overwrite a different existing file.

WAS THIS USEFUL?  

CopilotKit OpenBot

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

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

> OpenBot's headline addition is a Composio broker integration that opens Bot tool access to hundreds of SaaS apps, alongside a wave of hardening: named-field request validation across a dozen surfaces, escalation-failure audit logging, tool-grant cleanup, and stricter container isolation.

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

01 **Composio SaaS app broker integration** NEW 82

Adds Composio as a broker integration, providing a catalogue of connectable apps rather than a single connector. Setting `COMPOSIO_API_KEY` brokers a Bot's access to hundreds of SaaS apps via Composio, with per-person account connections, tool grants, and gateway-recorded calls; the Plugins page surfaces availability under More apps when unset.

Enable Composio so Bots can be granted tools from hundreds of connected SaaS apps without any per-app integration work.

env

```
COMPOSIO_API_KEY=<your-composio-key>
```

— Names env var and UI location, mechanism explained, but no numeric limits v0.0.12

02 **Named-field 400 validation across write/query surfaces**

IMPROVED 67

Refuses malformed requests with a named-field 400 response across the plugin server, tool-call endpoints, admin people search, credential input, skill tools, grant IDs, route IDs on routines, host-access, agents, channels, routing text length, routine dispatch, page-frame params, and the runtime env and model content surfaces.

— Enumerates every hardened surface but no sample request/response v0.0.12

03 **Hardened container OS isolation** IMPROVED 66

Runs the API and its migrations under a dedicated OS account in the all-in-one image, restricts browser environment variables to only what it needs, and closes `/run/s6/container_environment` to the Bot shell.

— Names exact restricted path but no reader-facing steps v0.0.12

THINNER COVERAGE BELOW

04 **Escalation-failure audit logging** NEW 53

Adds an `agent.escalation_failed` audit row carrying the actual error thrown when a person-routing call fails, and tells the Bot it could not reach anyone rather than surfacing the raw error.

— Names the audit field but gives no way to query or view it v0.0.12

05 **Tool grant cleanup on uninstall and upgrade** IMPROVED 42

Removes a skill's tool grants atomically on uninstall and drops orphaned grants left by prior uninstalls on upgrade.

— Describes behaviour but no command or config to trigger it v0.0.12

WAS THIS USEFUL? 👍 👎

Vercel AI SDK

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

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

> Vercel AI SDK's ai@7.0.102 release introduces experimental Live/Realtime voice capabilities, including an OpenAI Live provider, a browser-direct WebRTC path, and a new client-delegation-first WebSocket runtime, alongside breaking changes that strip out legacy realtime APIs.

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

01 **Breaking changes to realtime runtime APIs** BREAKING

80

The deferred managed Responses coordinator, `autoContinueTools` option, backend usage state, and Live tool-result aliases have been removed from the realtime WebSocket runtime. `sendTextMessage` and `addToolOutput` now reject continuous sessions, with server-confirmed provider delegation and turn-based audio-delta on the continuous profile failing explicitly, and `maxPlaybackBufferSeconds` is now supported only for continuous PCM sessions.

— Names every removed/restricted API for migration ai@7.0.102

02 **OpenAI Live provider via experimental_realtime** NEW

78

Adds `openai.experimental_realtime` factory that creates an OpenAI Live provider, routing known Live model IDs to Live sessions over server WebSocket connections. Includes client delegation, continuous audio and transcript events, microphone mute controls, graceful session close, and cumulative voice usage tracking.

— Names the exact factory and behaviours but no runnable snippet ai@7.0.102

03 **Browser-direct WebRTC for Live conversations** NEW 68

Adds an optional browser-direct WebRTC path for experimental client-delegated Live conversations, exchanging SDP through an application endpoint via `api.session`, with server-owned data-channel permissions and disconnect recovery.

— Names the config surface `api.session` but not full setup steps ai@7.0.102

04 **Client-delegation-first realtime WebSocket runtime** NEW

64

Adds a client-delegation-first realtime WebSocket runtime with session lifecycle management, bounded event queues, command acknowledgement tracking, transcripts,

context append, reversible capture, and a recoverable live-edge buffer policy for continuous PCM16 audio-chunk playback over an application relay.

— Rich mechanism detail but no named flag or entry point `ai@7.0.102`

05 **Live session reliability and error reporting** IMPROVED

60

Enforces `128 KiB` frame and combined buffered-send budgets for Live continuous sessions using actual encoded wire bytes, with oversized manual operations remaining recoverable and startup/automatic audio failures closing the session. Also reports abnormal realtime WebSocket close codes and reasons through the session error path.

— Gives a concrete size limit but no user-facing control `ai@7.0.102`

THINNER COVERAGE BELOW

06 **Realtime v4 spec extensions** IMPROVED

44

Extends the realtime v4 specification with optional server WebSocket configuration, per-connection raw-event parsers, and model-wide startup/finalization capabilities.

— Names spec additions but no concrete API surface `ai@7.0.102`

└──▷ **BREAKING ON UPGRADE**

! The deferred managed Responses coordinator, `autoContinueTools` option, backend usage state, and Live tool-result aliases have been removed from the realtime WebSocket runtime.

! `sendTextMessage` and `addToolOutput` now reject continuous sessions; server-confirmed provider delegation and turn-based audio-delta on the continuous profile fail explicitly.

! `maxPlaybackBufferSeconds` is supported only for continuous PCM sessions.

WAS THIS USEFUL?  

Mastra

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

Mastra is a TypeScript framework for AI applications and agents, providing workflows, agent memory, RAG, tool calling, and evals.

> Mastra 1.67 ships a Studio Workflow Builder backend, the @mastra/connect package for turning platform integrations into agent tools, per-call background execution dispositions with completion signals, and transactional thread ownership transfer, alongside breaking changes to thread event subscriptions and Archil filesystem's grep method.

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

01 Studio Workflow Builder backend NEW

90

Adds a Studio Workflow Builder backend via the `workflowBuilder` editor option on `MastraEditor`, exposing `GET /editor/workflow-builder/settings` and `POST /editor/workflow-builder/stream` endpoints gated by `stored-workflows:read` and `stored-workflows:write` permissions.

Enable the Studio Workflow Builder so an editor-owned agent can author and persist workflow definitions through the new server endpoints.

typescript

```typescript
import { Mastra } from '@mastra/core';
import { MastraEditor } from '@mastra/editor';

const mastra = new Mastra({
  editor: new MastraEditor({
    workflowBuilder: {
      enabled: true,
      model: 'openai/gpt-5.5',
      lastMessages: 100,
    },
  }),
});
```

— Named endpoints, permissions and runnable config example. `@mastra/core@1.67.0`

02 Background execution dispositions and completion signals NEW

90

Adds per-call background execution dispositions (`foreground` , `deferred` , `awaited`) to tools via `_background.disposition` in the call payload, plus a `defaultDisposition` field in the tool's `background` config block, and createBackgroundWorkSignalProcessor() from `@mastra/core/processors` for caller-

scoped background work completion signals (`work-completed` , `work-failed`) in agent streams.

Attach a background-work signal processor to an agent stream so your caller can react when deferred tool work completes or fails.

typescript

```typescript
import { Agent } from '@mastra/core/agent';
import { createBackgroundWorkSignalProcessor } from
'@mastra/core/processors';

const agent = new Agent({
  // ...existing agent config...
  inputProcessors: [createBackgroundWorkSignalProcessor()],
});

const stream = await agent.stream('Research distributed
systems');

for await (const part of stream.fullStream) {
  if (
    part.type === 'data-signal' &&
    (part.data.tagName === 'work-completed' || part.data.tagName
=== 'work-failed')
  ) {
    console.log('Background signal:', part.data.tagName);
  }
}
```

— Named fields, signal types and a runnable stream example. `@mastra/core@1.67.0`

03

hideSignals filtering across streaming and memory recall

NEW

80

Adds a `hideSignals` parameter to agent.stream(), agent.subscribeToThread(), agent.generate(), agent.resumeGenerate(), and the memory.recall() contract, accepting `true` , `false` / `[]` , or an array of signal type strings (e.g. `['reactive', 'system-reminder']`) to filter which streamed or stored signal messages callers receive, superseding the deprecated `includeSystemReminders` option.

— Names every affected method and accepted values, no code example. `@mastra/core@1.67.0`

04 **Transactional thread ownership transfer** NEW 80

Adds Memory.updateThreadResourceId({ threadId, resourceId }) and the server route `POST /memory/threads/:threadId/transfer` for transactional thread ownership reassignment, with MemoryThread.transfer() exposed in `client-js`.

— Named endpoint, method and client wrapper, no example shown. `@mastra/core@1.67.0`

05 **Strict error strategy for guardrail processors** NEW 80

Adds `errorStrategy: 'strict'` to model-backed guardrail processors (e.g. `PromptInjectionDetector`) so a failed internal model call trips a hard stop instead of failing open.

Harden a prompt-injection guardrail so any internal model failure blocks content rather than silently allowing it through.

typescript

```typescript
import { PromptInjectionDetector } from '@mastra/core';

const detector = new PromptInjectionDetector({
  model: 'openrouter/openai/gpt-oss-safeguard-20b',
  errorStrategy: 'strict',
});
```

— Named flag with a runnable configuration example. `@mastra/core@1.67.0`

06 **Experimental cross-agent communication tools** NEW 75

Adds experimental cross-agent communication tools (`agent_connections_list` , `agent_connect` , `agent_signal_send` , `agent_disconnect`) enabled via `crossAgentSignals: true` in createMastraCode() or the Experimental cross-agent communication toggle in `/settings` .

— Names all four tools and the exact enable path. `@mastra/core@1.67.0`

07 **@mastra/connect integration tools package** NEW 70

Adds the `@mastra/connect` package that turns Mastra Platform integration connections into agent tools with credential injection, token refresh, live tool discovery, and optional allowlisting via the `integrations` option.

— Named package and option, but no usage example given. `@mastra/core@1.67.0`

08 **subscribeThreadEvents replaces subscribeQueuedMessages** 70

BREAKING

Replaces subscribeQueuedMessages({ resourceId, threadId }, listener) with subscribeThreadEvents({ resourceId, threadId }, listener) — the listener now receives typed events such as `queue-count-changed`.

— Names old and new methods and an event type. `@mastra/core@1.67.0`

09 **Subagent result reference reuse in delegation** NEW

65

Adds `delegation.enableResultReferences` option to enable subagent result reuse via `[ref: <id>]` markers and a `contextFromRefs` input on delegation tools.

— Named option and marker syntax, no example provided. `@mastra/core@1.67.0`

10 **validateToolInput schema validation export** NEW 65

Adds `validateToolInput` export from `@mastra/core/tools` for validating values against a tool's input schema, returning `{ data, error }` (a `ValidationError` on mismatch).

— Named export and return shape, no usage example. `@mastra/core@1.67.0`

11 **Bulk notification status updates** NEW 60

Adds NotificationsStorage.updateNotificationsStatus({ threadId, ids, status }) to set a single status on many notifications in one write.

— Named method and signature, thin single-line addition. `@mastra/core@1.67.0`

THINNER COVERAGE BELOW

12 **hookDurationMs observability attribute** NEW 50

Adds `hookDurationMs` attribute to output stream processor observability spans, representing total time spent inside `processOutputStream` across all chunks.

— Named attribute and scope, no usage example. `@mastra/core@1.67.0`

13 **ArchilFilesystem.grep renamed to diskGrep** BREAKING

50

Renames ArchilFilesystem.grep() in `@mastra/archil` to diskGrep().

— Named rename, minimal surrounding context. `@mastra/core@1.67.0`

14 **Cross-trace thread identity queries for observability adapters**

NEW 25

Adds a storage contract for observability adapters to query thread identities using cross-trace predicates.

— Vague description with no named method or signature. `@mastra/core@1.67.0`

└──▷ **BREAKING ON UPGRADE**

! Replace subscribeQueuedMessages({ resourceId, threadId }, listener) with subscribeThreadEvents({ resourceId, threadId }, listener) — the listener now receives typed events such as `queue-count-changed`.

! ArchilFilesystem.grep() in `@mastra/archil` is renamed to diskGrep().

WAS THIS USEFUL?  

EXXETA exxperts

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗

RANK

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

> exxperts v0.12.0 focuses entirely on distribution, shipping signed desktop installers and self-contained server archives so the framework can be installed without Node, npm, or Git.

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

01 Signed desktop apps and portable server archives NEW

80

Adds signed, notarized desktop app builds (window + tray) for Windows x64 and macOS Apple Silicon that show no unknown-publisher prompt on first launch, plus prebuilt server archives for the `exxperts` command on Windows x64, macOS Apple Silicon, and Linux x64 bundling a vendored Node runtime so no Node, npm, or Git installation is required on the target machine. Versionless download aliases such as `exxperts-desktop-mac-arm64.dmg` and `exxperts-setup-win-x64.exe` are also added so `releases/latest/download` links always resolve to the newest release.

Verify the integrity of a downloaded archive before unpacking — use `--ignore-missing` so only the file you actually downloaded is checked.

```
$ shasum -a 256 --ignore-missing -c SHA256SUMS.txt
```

— Names exact platforms, filenames, and a verification command; mechanism partly explained

v0.12.0

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

Together AI

Feedback ⓘ 1 RELEASE · SEEN 2026-09-16 NOTES ↗ RANK

Run, train, and serve open-source AI models on Together AI.

> Together AI shipped a full canary rollout management API for endpoints, added automatic metric-gate pausing for regressing rollouts, and introduced reserved and on-demand GPU cluster creation.

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

01 **Canary rollout lifecycle API for endpoints** NEW 95

Adds a full rollout management surface: the List Rollouts endpoint gains a `filter=ROLLOUT_FILTER_ACTIVE` query parameter to return only the active rollout; a Create Rollout API creates a rollout in the pending state without shifting traffic so strategy and metric gates can be reviewed before a separate Start request; a Start Rollout API begins the configured traffic-shifting workflow; a Get Rollout API retrieves a rollout's strategy, lifecycle state, current traffic percentage, step history, and metric-gate results; a Pause Rollout API requests a running or stabilizing rollout to pause, accepts an optional reason, and returns the PAUSING snapshot; a Resume Rollout API resumes a pausing, paused, or system-paused rollout from its current step and traffic split; a Promote Rollout API completes a rollout immediately by sending all live traffic to the target deployment; a Cancel Rollout API freezes the current traffic split into standing weights and brings the rollout to CANCELED state; and a Delete Rollout API removes a rollout record after it has been aborted or completed.

List only the currently active rollout on an endpoint to quickly confirm live traffic state without wading through historical records.

```
$ curl -X GET
'https://api.together.ai/v1/endpoints/<endpoint_id>/rollouts?
filter=ROLLOUT_FILTER_ACTIVE' \
  -H 'Authorization: Bearer <TOGETHER_API_KEY>'
```

— Every rollout endpoint named with states and a runnable curl example snapshot-20260916

THINNER COVERAGE BELOW

02 **Metric-gate auto-pause for rollouts** NEW 45

Metric-gate support for canary rollouts automatically pauses a rollout when the new deployment regresses, feeding into the Pause/Resume/Promote rollout lifecycle.

— Mechanism described but no thresholds or config named snapshot-20260916

03 **Reserved and on-demand GPU clusters** NEW 25

Supports reserved and on-demand cluster creation with connectivity to workloads.

— Thin one-line description with no commands or limits `snapshot-20260916`

WAS THIS USEFUL? 👍 👎

FlashAttention

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗

RANK

FlashAttention implements fast, memory-efficient exact attention kernels used by most current transformer training and inference stacks.

> FlashAttention's FA4 beta31 release adds Attention Sink support to Flash MLA on SM100 GPUs and extends the SM100 HD256 2-CTA decoder path with new sequence-length and paged-KV handling.

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

01 **SM100 HD256 2-CTA decoder path enhancements** NEW 70

Adds `seqused_q` / `seqused_k` support and the decoder paged-KV shape to the SM100 HD256 2-CTA path.

— Names exact parameters and path added `fa4-v4.0.0.beta31`

THINNER COVERAGE BELOW

02 **Attention Sink support in Flash MLA on SM100** NEW 40

Adds Attention Sink support to Flash MLA kernels running on SM100 (Blackwell) GPUs.

— Names GPU target and mechanism but no usage detail `fa4-v4.0.0.beta31`

WAS THIS USEFUL? 👍 👎

Eigen Labs Darkbloom

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ CODE ↗ RANK

Private Inference Network on Idle Macs

> Darkbloom v0.9.4 upgrades provider attestation to App Attest protocol 3 with SDK 27 CodeDirectory measurements, and extends its private dashboard and Datadog telemetry to track build approval, enrollment recovery and warm-pool planning.

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

01 SDK 27 CodeDirectory measurements for build approval NEW

75

Captures SDK 27 Apple-signed CodeDirectory measurements and requires an exact qualified binary/code-hash pair for prospective build approval, with release builds selecting SDK 27 / Swift 6.4 and recording the full CodeDirectory SHA-256.

— Names exact SDK, Swift version and hash mechanism v0.9.4

02 Attestation dashboard with cohort readiness tracking NEW

70

Records versioned prospective authorization outcomes, revocation, current-connection freshness, qualified builds and receipt/risk readiness, with the private dashboard showing exact cohort denominators and blockers for a later MDM retirement.

— Names dashboard destination and tracked fields v0.9.4

03 Coordinator telemetry to Datadog for warm-pool planning NEW

70

Forwards the latest distinct per-model warm-pool planning snapshot through the coordinator telemetry emitter so Datadog can show target sizing, measured demand, candidate availability and blocker counts.

— Names Datadog integration and specific tracked metrics v0.9.4

04 App Attest protocol 3 identity binding NEW

60

Negotiates App Attest protocol 3 to bind app-measured machine model, RAM, CPU/GPU counts and the existing verification key to the signed transcript, preserving old transcripts and cached enrollment recovery across upgrades.

— Describes protocol upgrade mechanism; no user-facing action v0.9.4

THINNER COVERAGE BELOW

05 **Machine identity association via App Attest assertions** NEW

40

Associates machine identities through fresh account-bound App Attest assertions, including reconnects after legacy-key rotation.

— Brief mechanism, no operational surface named v0.9.4

06 **Cached enrollment receipt recovery** IMPROVED

40

Recovers cached enrollment receipts through a separately validated renewal path, retaining original failure records and reconciling interrupted evidence without advancing counters.

— Explains renewal path but names no operational surface v0.9.4

07 **Retry handling for transient shadow failures** IMPROVED

30

Retries transient shadow failures, including temporarily saturated control-lane sends, without disconnecting serving providers.

— Thin description, no named surface v0.9.4

WAS THIS USEFUL?  

Anthropic TypeScript SDK

Feedback ⓘ

1 RELEASE · 2026-09-15 NOTES ↗ RANK

The Anthropic TypeScript SDK provides typed JavaScript and TypeScript clients for Claude APIs, streaming, and tool use.

> The Anthropic TypeScript SDK's latest release adds compaction controls to manage context windows in long agentic loops, lets the web fetch tool retrieve caller-specified URLs, and rounds out several smaller API and bundling improvements.

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

01 Compaction control for agentic tool-use loops NEW

93

Adds a `compaction` parameter and signed compaction blocks (beta) to control context window summarization in long-running agentic loops, configurable via `compactionControl` with `enabled` and `contextTokenThreshold` fields. This lets callers set a token threshold at which the SDK automatically summarizes context to avoid hitting model limits.

Prevent a long agentic tool-use loop from hitting the context limit by enabling auto-compaction with a token threshold.

typescript

```
const runner = client.beta.messages.toolRunner({
  model: 'claude-sonnet-5',
  max_tokens: 4096,
  tools: [search, done],
  messages: [{ role: 'user', content: 'Analyze all quarterly
reports.' }],
  compactionControl: {
    enabled: true,
    contextTokenThreshold: 3000,
  },
});

for await (const message of runner) {
  console.log(`Input tokens: ${message.usage.input_tokens}`);
}
```

— Full mechanism, config fields, and a runnable code example. sdk-v0.126.0

THINNER COVERAGE BELOW

02 **`url_sources` for the web fetch tool** NEW 55

Adds `url_sources` to the web fetch tool, allowing callers to supply specific URLs for the tool to retrieve rather than relying on search discovery.

– Named field and behavior, but no usage example. sdk-v0.126.0

03 **`workspace_id` scoping for user profile methods** NEW 55

Adds a `workspace_id` parameter to user profiles methods, scoping profile operations to a specific workspace.

– Names the exact parameter and its scoping effect. sdk-v0.126.0

04 **`thinking_mismatch_allowed` in `input_transformations`** NEW 45

Adds `thinking_mismatch_allowed` entries to `input_transformations` (beta) for finer control over thinking-mode request transformations.

– Named beta field but no example of its use. sdk-v0.126.0

05 **Tree-shaking support for bundlers** IMPROVED 45

Adds `'sideEffects': false` and pure-class annotations so bundlers can tree-shake unused SDK modules, reducing bundle sizes in browser and edge deployments.

– Names the package.json flag and effect, no numbers given. sdk-v0.126.0

06 **Enum types for workspace data-residency `geo` fields** IMPROVED 43

Adds enum types for workspace data-residency `geo` fields, replacing free-form strings with validated values.

– Names the field but no values or usage detail. sdk-v0.126.0

07 **Auto mode tool permissions for Managed Agents** NEW 30

Adds auto mode tool permissions for Managed Agents, enabling automatic permission handling for agent toolsets.

– Described only in one generic sentence, no mechanism or surface. sdk-v0.126.0

WAS THIS USEFUL?  

OpenAI TypeScript SDK

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

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

> OpenAI TypeScript SDK v7.16.0 adds per-iterator event limits for WebSocket streaming, giving finer control over how many events an async iterator will consume.

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

01 **Per-iterator event limits for WebSocket streams** NEW

35

Each async iterator over a WebSocket stream can now cap the number of incoming events it will consume, letting callers bound event consumption on a per-iterator basis.

— Describes mechanism but no exact API name or option key given v7.16.0

WAS THIS USEFUL? 👍 👎

Hugging Face PEFT

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗

RANK

PEFT implements parameter-efficient fine-tuning methods including LoRA, QLoRA, prefix tuning, and adapters for large pretrained models.

> PEFT v0.21.0 adds four new fine-tuning methods (Riemannian LoRA optimizer, KaSA, Super-Tuning, ShadowPEFT), speeds up MoE LoRA delta computation, adds torchao v0.18.0 tensor support, cuts OSF memory and training cost, and tightens adapter lifecycle validation.

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

01 Adapter lifecycle and validation improvements IMPROVED

80

Prompt learning methods now support deleting an adapter, and for adaption prompt, loading or adding a second adapter no longer automatically activates it. `inject_adapter_in_model` now raises an error when called with PEFT methods that have shared state requiring a parent model (such as VeRA). `add_adapter` now raises a `peft.PeftError` when the new adapter targets non-existing layers instead of silently adapting nothing, and `peft_model.targeted_module_names` and `peft_model.targeted_parameter_names` no longer contain duplicates. Deleting an adapter now raises an error if the adapter is currently merged, since deletion would not remove its effect.

— Names every function and attribute changed v0.21.0

02 Super-Tuning and Supra sparse tuning method NEW

72

Adds Super-Tuning as a new PEFT method that uses calibration-free magnitude scoring with `BottomK` selection to allocate a sparse tuning budget, with an optional Supra variant that layers LoRA on top of the sparse update.

— Names selection strategy and variant but no config example v0.21.0

03 OSF memory and training time reduction IMPROVED

70

Reduces Orthogonal Subspace Fine-tuning (`OSF`) memory footprint by 22% and training time by 46% on the MetaMathQA benchmark while keeping task accuracy and forgetting stable.

— Concrete benchmark numbers but no mechanism named v0.21.0

04 KaSA LoRA variant NEW

67

Adds `KaSA` (Knowledge-Aware Singular-Value Adaption) as a new LoRA variant that identifies knowledge-aware singular values from base weights, removes noisy

knowledge, and trains only the most task-relevant singular values.

— Named method with mechanism but no usage snippet v0.21.0

05 **ShadowPEFT method** NEW 65

Adds `ShadowPEFT` as a new PEFT method that runs a small shadow network alongside the frozen base model, sharing adaptation state across all decoder layers instead of per-layer low-rank updates.

— Mechanism described, no usage instructions given v0.21.0

06 **Riemannian-preconditioned LoRA optimizer** NEW 60

Adds a Riemannian-preconditioned LoRA optimizer that applies Riemannian pre-conditioning on gradients to improve feature learning and eliminate the need for separate learning-rate tuning for LoRA's A and B matrices.

— Mechanism explained but no config key or flag given v0.21.0

THINNER COVERAGE BELOW

07 **Faster MoE LoRA delta computation** IMPROVED 55

Speeds up MoE LoRA delta computation for the single-active-adapter case by folding weight and delta computation into a single `baddbmm` operation.

— Names the operation but no benchmark numbers v0.21.0

08 **torchao v0.18.0 tensor subclass support** NEW 50

Adds support for torchao v0.18.0, including its newer `Int8Tensor`, `Float8Tensor`, and related tensor subclasses.

— Names version and tensor types, no usage detail v0.21.0

WAS THIS USEFUL?  

XGrammar

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

XGrammar constrains language-model output to grammars and schemas with optimized token-mask generation.

> XGrammar v0.2.7 adds structural tag support for DeepSeek V4.1 and MiniMax M3 tool-call formats, a new `parallel_tool_calls` parameter for builtin structural tags, and canonical recursive Any support for Cohere XML.

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

01 Structural tags for DeepSeek V4.1 and MiniMax M3 NEW

60

Adds the `deepseek_v4_1` structural tag and `deepseek_v4_1_xml` schema style for spaced DSML tags used by DeepSeek-V4.1-Flash, aligned with the official encoder and tokenizer, and the `minimax_m3_xml` structural tag for MiniMax M3 tool-call formatting. Both are exposed through Python, C++, and TypeScript.

— Names exact tag identifiers but no usage example given v0.2.7

THINNER COVERAGE BELOW

02 parallel_tool_calls parameter for structural tags NEW

40

Adds a `parallel_tool_calls` parameter to the builtin structural tags, available across Python, C++, and TypeScript.

— Names the parameter but not its behavior or values v0.2.7

03 Canonical recursive Any for Cohere XML NEW

25

Adds canonical recursive Any support for Cohere XML.

— Bare mention with no mechanism or usage detail v0.2.7

WAS THIS USEFUL? 👍 👎

Anthropic Python SDK

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

The Anthropic Python SDK provides typed Python clients for Claude APIs, streaming responses, and tool use.

> Anthropic Python SDK v1.6.0 adds explicit URL sourcing for the web fetch tool, beta context-window compaction controls, auto mode tool permissions for Managed Agents, async credential token providers, and workspace profile/data-residency refinements.

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

01 **URL sources for web fetch tool** NEW 65

Adds a `url_sources` parameter to the web fetch tool, letting callers supply explicit URLs for the agent to fetch instead of letting it choose which pages to retrieve.

Pass `url_sources` to the web fetch tool to instruct the agent to retrieve specific URLs rather than letting it choose.

python

```
import os
from anthropic import Anthropic

client = Anthropic(api_key=os.environ.get("ANTHROPIC_API_KEY"))

response = client.messages.create(
    model="claude-opus-5",
    max_tokens=1024,
    tools=[
        {
            "type": "web_fetch",
            "url_sources": [
                "https://example.com/advisory.html",
                "https://nvd.nist.gov/vuln/detail/CVE-2024-1234",
            ],
        }
    ],
    messages=[{"role": "user", "content": "Summarise the linked
advisories."}],
)
print(response.content)
```

— Named parameter with runnable example, but no mechanism detail v1.6.0

THINNER COVERAGE BELOW

02 **Beta compaction parameter for context windows** NEW

55

Adds a `compaction` parameter and signed compaction blocks to the API (beta), enabling context-window compaction control.

— Named beta parameter but no usage example or mechanism v1.6.0

03 **Thinking mismatch handling in input transformations** NEW

55

Adds `thinking_mismatch_allowed` entries to `input_transformations` (beta) for finer control over thinking-mode input handling.

— Names exact beta field but lacks usage example v1.6.0

04 **Workspace profile and data-residency enhancements** IMPROVED

55

Adds a `workspace_id` parameter to user profiles methods for workspace-scoped profile queries, and introduces enum types for workspace data-residency geo fields, replacing raw strings with validated values.

— Names both surfaces but no example of use v1.6.0

05 **Auto mode tool permissions for Managed Agents** NEW

35

Adds auto mode tool permissions for Managed Agents, reducing manual permission configuration.

— States benefit but no config surface or mechanism named v1.6.0

06 **Async credential token providers** NEW

35

The client now supports async credential token providers, enabling non-blocking token refresh workflows.

— Describes capability but no interface or code path given v1.6.0

WAS THIS USEFUL? 👍 👎

Baseten

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ RANK

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

> Baseten's CLI reached 1.0.0 general availability, marking a stable command surface for the tool.

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

01 **Baseten CLI 1.0.0 general availability** NEW 20

Baseten CLI 1.0.0 is now generally available with a stable command surface.

— No commands, flags, or changes named beyond version bump Baseten CLI 1.0.0

WAS THIS USEFUL? 👍 👎

Snowflake Cortex AI

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

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

> Snowflake introduced Cortex AI Gateway (Preview), a single control plane for calling models via OpenAI- and Anthropic-compatible endpoints, tracking and capping spend, tracing requests, and controlling who can reach it.

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

01 Usage tracking, quotas and budgets in AI Gateway NEW

80

Adds the `AI_GATEWAY_USAGE_HISTORY` view in the `ACCOUNT_USAGE` schema, recording gateway, user, model and token consumption per request; per-user quotas that block an individual user once they hit a defined token/credit limit; support for adding the gateway as a shared resource to a custom budget so credit consumption can be tracked by team or cost center; and a Snowsight dashboard under AI & ML » Cortex AI Gateway showing request/response counts, token counts and latency across models without writing SQL.

Inspect gateway activity visually — request counts, latency, and model-level token spend — without writing SQL.

> 📍 In Snowsight, go to AI & ML › Cortex AI Gateway to view the built-in dashboard showing request/response counts, token counts, and latency broken down by model and user.

— Names the view, quota limit, budget mechanism and dashboard path

Sep 15, 2026: Cortex AI Gateway (Previe…

02 USAGE privilege controls on the gateway object NEW

80

The `SNOWFLAKE` gateway object supports a `USAGE` privilege, granted to `PUBLIC` by default; `ACCOUNTADMIN` can revoke and re-grant it to restrict gateway traffic to specific roles, e.g. `REVOKE USAGE ON GATEWAY SNOWFLAKE FROM ROLE PUBLIC; GRANT USAGE ON GATEWAY SNOWFLAKE TO ROLE AI_USERS;`.

Lock down the gateway so only the `AI_USERS` role can reach it, removing the default open-to-everyone access.

```
$ REVOKE USAGE ON GATEWAY SNOWFLAKE FROM ROLE PUBLIC;
GRANT USAGE ON GATEWAY SNOWFLAKE TO ROLE AI_USERS;
```

— Runnable SQL commands given for locking down access

Sep 15, 2026: Cortex AI Gateway (Previe…

THINNER COVERAGE BELOW

03 **Unified inference endpoint in Cortex AI Gateway** NEW

55

The gateway adds a single inference endpoint that accepts both the OpenAI Chat Completions and Anthropic Messages API request formats, letting official SDKs or coding agents such as OpenCode call Snowflake-governed models through one interface.

— Names two API formats and OpenCode but no endpoint URL or sample request

Sep 15, 2026: Cortex AI Gateway (Previe…

04 **Per-gateway trace tables for observability** NEW 55

Adds per-gateway trace tables exposing spans for any request, including which models were called, step timing, token counts and errors.

— Names trace fields but gives no query example Sep 15, 2026: Cortex AI Gateway (Previe…

05 **Optional prompt/response payload logging** NEW 35

The gateway supports optional payload logging of prompts and responses, off by default, while metadata-level recording of every request is always on.

— No config key or command shown to enable it Sep 15, 2026: Cortex AI Gateway (Previe…

WAS THIS USEFUL? 👍 👎

Ollama

Feedback ⓘ 1 RELEASE · 2026-09-14 NOTES ↗ CODE ↗ RANK

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

> Ollama v0.34.1 makes `/api` /tags roughly 10x faster, tightens repeat-token error handling, and removes in-tree safetensor-to-GGUF conversion in favor of external llama.cpp tooling, while also deprecating the typical_p sampling parameter for new models.

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

01 **Faster /api/tags with per-blob metadata caching** IMPROVED

67

`/api/tags` is now roughly 10× faster on large model libraries, dropping cold response time from 3.1s to 294ms, and reports model capabilities more consistently by caching GGUF metadata per blob.

— Named endpoint and concrete before/after timings given. v0.34.1

02 **In-tree safetensor-to-GGUF conversion removed from ollama create**

67

BREAKING

GGUF model creation via `ollama create` no longer supports in-tree safetensor conversion or server-side quantization; users must convert and quantize safetensors to GGUF with llama.cpp tooling before importing.

— Names exact command and gives migration path. v0.34.1

THINNER COVERAGE BELOW

03 **typical_p sampling parameter deprecated** DEPRECATED 57

The `typical_p` sampling parameter is deprecated: it can no longer be set when creating new models, though existing GGUF models that already use it retain support.

— Names the parameter and exact scope of removal. v0.34.1

04 **Stricter runaway repeat-token detection** IMPROVED 53

Runaway repeat-token detection now requires 100 repeated tokens before aborting, and returns an HTTP 500 error with a descriptive message instead of silently returning an incomplete result.

— Mechanism and threshold given but nothing for reader to invoke. v0.34.1

05 **MLX safetensors import graduates from experimental** IMPROVED

23

MLX safetensors import is promoted out of experimental status, per the release summary, though no further mechanism or scope is documented.

— Only mentioned in summary; no detail on behavior or usage. v0.34.1

└──▷ **BREAKING ON UPGRADE**

! GGUF model creation via `ollama create` no longer supports in-tree safetensor conversion or server-side quantization — use llama.cpp tooling to convert and quantize safetensors to GGUF before importing.

! `typical_p` can no longer be set when creating new models.

WAS THIS USEFUL? 👍 👎

OpenRouter

Feedback ⓘ 1 RELEASE · 2026-09-08 NOTES ↗ RANK

OpenRouter is a routing gateway that exposes hundreds of models from many providers behind one OpenAI-compatible API with failover and unified billing.

> OpenRouter introduced a hosted Linux shell tool and companion Files API, letting any model on the platform execute commands and move files through isolated containers.

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

01 Hosted shell tool and Files API for model execution NEW

70

OpenRouter added a shell server tool that gives any model a hosted, isolated Linux container, returning stdout, stderr, and exit codes from commands the model runs. A companion Files API moves files into and out of that container, including downloading files written during a run. Both are available in beta on the Responses and Messages APIs.

— Names container behavior and APIs but no exact endpoint paths or flags.

snapshot-20260916

WAS THIS USEFUL? 👍 👎

RunPod

Feedback ⓘ 1 RELEASE · 2026-09-01 NOTES ↗ RANK

AI infrastructure with on-demand GPUs and serverless compute. Run training, inference, and batch workloads on the cloud with Runpod.

> RunPod introduces Global Volumes, a new beta storage type providing elastic, region-independent storage that any Pod can mount without copying data between data centers.

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

01 **Global Volumes storage type for Pods** NEW 93

RunPod adds Global volume as a new storage type alongside network volumes — elastic, region-independent, object-backed storage that persists across Pod stop/delete and grows automatically with no upfront capacity provisioning. It mounts by default at `/workspace`; if a network volume is also attached, the deploy flow automatically remounts the global volume to `/workspace-global` (or `/workspace-2` if the mount path was manually edited). Each Pod supports at most one global volume and one network volume simultaneously, and the storage picker tags global volumes with a Global badge. To attach one during deployment, set the Persistent storage type to Storage volume and pick it from the Global volumes dropdown. Use cases include model serving/inference (loading weights, tokenizer files, configs), shared LoRA adapters/embedding indices across multiple Pods, and read-only config distribution across a Pod fleet. Data can be migrated from a network volume with `rsync --archive --verbose --human-readable /workspace/ /workspace-global/`, and mounting can be verified with `ls /workspace-global`. Because it's object-backed, it does not support file locking, atomic rename, hard links, file permission bits, or concurrent writes across Pods; accounts with a $0 balance have their global volume flagged and permanently deleted after 15 days.

Migrate existing model weights from a network volume to a global volume so they can be read by Pods in any region.

```
$ rsync --archive --verbose --human-readable /workspace/
/workspace-global/
```

Verify that your global volume is correctly mounted inside a running Pod before kicking off an inference workload.

```
$ ls /workspace-global
```

Create and attach a global volume to a new Pod through the RunPod console UI.

📍 1. In the left nav, go to Storage. 2. Click `+ New volume`, select `Global volume` under Storage type, enter a Name, then click `Create global volume`. 3. Go to Pods and click `+ Deploy`. Select a template and GPU. 4. Scroll to the Storage section, click `+ Add volume` under Persistent storage, and select your global volume (shown with a `Global` badge). 5. Click `Deploy Pod`.

Migrate model weights from a network volume to a newly attached Global Volume so the data becomes region-independent and readable by any Pod.

```
$ rsync --archive --verbose --human-readable /workspace/
/workspace-global/
```

Verify a Global Volume is mounted correctly inside a running Pod before serving inference workloads.

```
$ ls /workspace-global
```

Create and attach a Global Volume to a new Pod through the RunPod console so model weights are available at startup across any region.

📍 1. In the left nav, go to Storage. 2. Click + New volume. 3. Under Storage type, select Global volume, enter a Name, and click Create global volume. 4. Go to Pods and click + Deploy. 5. Select a template and GPU, then scroll to the Storage section. 6. Under Persistent storage, set the type to Storage volume. 7. Click the Storage volume dropdown and select your volume from the Global volumes section. 8. Click Deploy Pod.

— Full mechanism, named paths, flags, limits and deploy steps given.

New ReleaseGlobal Volumes - Beta

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

Comfy Org ComfyUI

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗

RANK

ComfyUI is a node-graph interface, API, and backend for running diffusion models, chaining generation steps into reusable visual workflows.

> ComfyUI v0.36.0 introduces generic loop execution for iterative workflows, a Video Concatenate node, Marigold v2 depth estimation, and YuE2 music generation, alongside a broad wave of new partner-node integrations spanning OpenRouter, Gemini, Tripo, Bria, BFL, and Pruna.

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

01 **Partner node integrations expanded** NEW 75

The OpenRouter partner node adds Microsoft `mai-image-2.6` models and an auto aspect ratio option to the MAI image node. The Gemini text partner node adds Gemini 3.8 Flash and a `GeminiNodeV3`. The Tripo partner node is migrated to the v3 API and gains a Smart Segment node plus Tripo P2 text, image, and multiview-to-model nodes. The Bria partner node adds new image edit nodes and a Video Eraser node. The BFL partner node adds a Flux Video Edit node. The Pruna partner node adds text-to-video and image-to-video nodes powered by P-Video-2. Partner node API responses now expose estimated-duration headers to drive progress display in the UI.

— Enumerates every named partner, model, and node added this release. v0.36.0

02 **YuE2 music generation support** NEW 65

Adds support for the YuE2 music generation model, including an ABC-notation generation node with expanded controls and an increased maximum song duration.

Generate music from an ABC-notation prompt using the Yue2 model with the expanded song-duration and generation controls.

📍 In the ComfyUI canvas, right-click › Add Node › YuE2 › Generate ABC, set your desired duration and style parameters, then connect the output to a Yue2 Audio Decode node and a Save Audio node.

— Names the model, node, and specific improvements with a usage example. v0.36.0

03 **Generic Loops for iterative graph execution** NEW 60

ComfyUI now implements Generic Loops, enabling iterative graph execution workflows where nodes between loop entry and exit boundaries repeat until a condition is met.

Heterogeneous lists now correctly carry every item through generic loop iterations, fixing a data-handling issue in looped workflows.

Build an iterative refinement workflow that repeatedly processes nodes until a condition is met, using the new Generic Loops capability.

> 📍 In the ComfyUI canvas, right-click › Add Node › Loop, then connect your processing nodes between the loop entry and exit boundary nodes to repeat execution for each iteration.

— Names the mechanism and a list-handling fix, with UI steps to build it. v0.36.0

THINNER COVERAGE BELOW

04 **Video Concatenate node** NEW 50

Adds a Video Concatenate node for joining multiple video clips together inside a workflow, connecting video outputs to a single combined result.

Stitch together multiple generated video clips into a single output without leaving ComfyUI.

> 📍 In the ComfyUI canvas, right-click › Add Node › Video Concatenate, then connect two or more video output nodes to its inputs and wire the result to a video preview or save node.

— Named node with a concrete usage path, but no deeper mechanism described. v0.36.0

05 **Linear color space option in ImageColorSpace** IMPROVED 50

Adds a `linear` color space option to the `ImageColorSpace` node.

— Names the exact node and new field value but no usage example. v0.36.0

06 **Marigold v2 depth-estimation support** NEW 25

Adds Marigold v2 support for depth-estimation workflows.

— Bare mention of support with no node names or usage detail. v0.36.0

07 **Model blueprints and subgraph reorganization** IMPROVED 25

Adds Marigold V2 and YuE2 model blueprints, adds new model blueprints generally, and reorganizes subgraph categories.

— Only vaguely describes new blueprints and reorganized categories. v0.36.0

WAS THIS USEFUL?

Mozilla AI llamafile

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

llamafile packages a model and its runtime into a single executable file that runs on macOS, Linux, and Windows without installation.

> llamafile 0.10.6 focuses on GPU acceleration, adding a Vulkan backend and enabling CUDA/HIP graph execution to reduce kernel launch overhead.

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

01 **CUDA/HIP graph execution flags** IMPROVED 70

llamafile enables the `-DGGML_CUDA_USE_GRAPHS` and `-DGGML_HIP_GRAPHS` build flags, turning on CUDA and HIP graph-based GPU execution to reduce kernel launch overhead during inference.

— Names exact build flags but no benchmark numbers. 0.10.6

THINNER COVERAGE BELOW

02 **Vulkan GPU backend support** NEW 40

Adds Vulkan toolchain support, enabling llamafile to run inference on Vulkan-capable GPUs.

— States the new backend but no setup or flag details. 0.10.6

WAS THIS USEFUL?

KoboldCpp

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗ CODE ↗

RANK

KoboldCpp runs GGUF language models locally as a single-file application with a built-in web UI and API.

> KoboldCpp v1.121 adds CPU-offload and metadata-analysis flags for model loading, a Docker LoRA-provisioning environment variable, and a batch of Stable Diffusion UI upgrades including a runtime LoRA selector and connection-loss recovery.

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

01 **CPU offload flag for FFN layers** NEW 80

Adds `--ffncpu` flag, behaving like llama.cpp's `--n-cpu-ffn`, as an alternative way to partially offload FFN layers to CPU, letting users keep feed-forward network layers off the GPU while other layers remain accelerated.

Run a model with partial GPU offload while keeping FFN layers on the CPU to reduce VRAM usage on constrained hardware.

```
$ koboldcpp --model model.gguf --gpulayers 20 --ffncpu
```

— named flag, comparison to llama.cpp equivalent, runnable command given v1.121

02 **KCPP_LORA env var for Docker LoRA provisioning** NEW 75

Adds `KCPP_LORA` environment variable to the Docker image for supplying comma-separated LoRA download URLs at container startup, enabling automatic fetching of LoRAs for reproducible deployments.

Pull LoRAs automatically into a KoboldCpp Docker container at startup — useful for reproducible cloud GPU deployments without baking weights into the image.

```
$ docker run -e
KCPP_LORA="https://example.com/lora1.safetensors,https://example.
com/lora2.safetensors" koboldai/koboldcpp
```

— named env var with a runnable docker run example v1.121

03 **SDUI runtime LoRA selector, connection recovery, and video limits**

NEW

75

The Stable Diffusion UI at `http://localhost:5001/sdui` gains a runtime LoRA selector for viewing, adding, and removing image generation LoRAs at runtime; a Poor connection mode to recover generated images after disconnection or an accidental browser close; a box for sending extra custom fields; a max video frame increase to 240 by default (480 with an extended toggle); and a display of elapsed generation time.

— concrete UI path and numeric limits, but only a UI path, no API v1.121

---

04 **Safetensor metadata analysis via --analyze** IMPROVED

65

Extends `--analyze` to inspect `.safetensor` file metadata in addition to `.gguf` files, letting users audit model metadata without loading weights into VRAM.

Inspect metadata of a `.safetensor` model file without loading it — useful for quickly auditing architecture, tensor names, or dtype before committing VRAM.

```
$ koboldcpp --analyze model.safetensors
```

— named flag and format with a runnable inspection command v1.121

THINNER COVERAGE BELOW

05 **Minimax H3 media reference support** NEW

40

Adds media reference support for Minimax H3, enabling attachment of audio clips and multiple reference images for video generation.

— names the model integration but no API or UI path given v1.121

---

06 **Video LoRA support in generation** NEW

25

Adds support for video LoRAs in image/video generation, extending the existing LoRA system beyond image models.

— one-line addition with no mechanism or usage path given v1.121

---

07 **xhigh reasoning effort level** NEW

15

Adds support for `xhigh` reasoning effort level.

— single-line addition with no parameter context or usage v1.121

---

08 **MusicUI rework** IMPROVED

10

Reworks the Music Generation UI (MusicUI) for a more intuitive user experience.

— vague rework with no specifics on what changed v1.121

WAS THIS USEFUL? 👍 👎

MLX Serve

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗ RANK

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

> mlx-serve v26.9.3 adds ANE offload for image/video/audio generation, idle model eviction, several speculative-decoding tuning flags, two new native model pulls, and a handful of breaking changes around KV-quant schemes and cost-table persistence.

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

01 **Native pulls for K2-Horizon and Spark-X2.5** NEW 88

`mlx-serve pull k2` runs K2-Horizon 7B natively with thinking at all three effort levels, JSON schema, tool calls, and a 512k context window. `mlx-serve pull spark` runs Spark-X2.5 1.7B/4B natively with thinking, tools, and a 1M context window.

— Exact pull commands with model sizes, capabilities and context limits. v26.9.3

02 **MTP speculative decoding tuning flags** NEW 85

New `--mtp-typical 0.2` and `--mtp-tokenv3 0.95` flags (also via Model Settings > MTP acceptance) trade speculation exactness for speed, giving 7 to 27% faster decode at temperature 1. A new `--mtp-head-kv-quant` flag stores Flash Next's speculative-head KV cache at the model's `--kv-quant` precision, saving ~1 GB at 1M tokens with no measurable acceptance loss; off by default.

— Three exact flags with concrete speed and memory tradeoffs. v26.9.3

03 **ANE offload for image/video/audio generation** NEW 83

New `--ane-image`, `--ane-video`, `--ane-audio` flags (also settable via Settings > Neural Engine) split each Krea, MiniMax-H3, or ACE-Step generation step between GPU and ANE, giving 1.3x on Krea, 1.33x on ACE-Step, and 1.22x per step on MiniMax-H3 on an M4 Max.

— Three named flags with per-model speedup numbers on a specific chip. v26.9.3

04 **GGUF speculation via own MTP head** NEW 79

GGUFs with their own MTP head can now be used for speculation on greedy and sampled requests via `--mtp` (on by default; opt out with `--no-ds4-mtp`), reaching 35 to 47 tok/s on the Flash Next Q2 pack on an M4 Max.

— Named flags, opt-out, and measured throughput on a specific pack. v26.9.3

05 **Concurrent request batching visibility** NEW 73

Whether a loaded model supports concurrent request batching is now exposed via Settings > Concurrent requests and the `/props` endpoint, e.g. `curl http://localhost:8080/props`.

Check whether the loaded model supports concurrent request batching and inspect server properties via the `/props` endpoint.

```
$ curl http://localhost:8080/props
```

— Runnable curl example against a named endpoint provided. v26.9.3

06 **OpenCode 2 dashboard plugin** NEW 70

`mlx-serve launch opencode2` installs an OpenCode 2 dashboard plugin showing tok/s, memory, model, and prefill progress; requires the server to be started with `--metrics`.

— Exact command and its dependency are given, mechanism is thin. v26.9.3

07 **Idle model eviction** NEW 67

New `--idle-evict-secs N` flag (also via Settings > Server > Unload idle models) automatically unloads models idle for N seconds; disabled by default.

— Exact flag and default given, but no mechanism beyond timeout. v26.9.3

08 **Wired memory margin flag for large Macs** NEW 66

New `--wired-margin-gib 6` flag frees ~2 GB on 96 GB Macs when combined with a raised `iogpu.wired_limit_mb`.

— Names the flag and companion config key with a concrete gain. v26.9.3

09 **Repetition-loop finish reason** IMPROVED 65

Replies cut short for repetition now return `finish_reason: "stop"` with `finish_details: {"type": "repetition_loop"}`, replacing the previous response that looked like a token-limit cutoff.

— Exact field values given with clear before/after behaviour. v26.9.3

10 **Speculative cost table persistence now opt-in** BREAKING 62

The speculative cost table is no longer persisted across server restarts by default; setting the new `MLX_SERVE_ROUND_COST_PERSIST=1` environment variable restores the previous persist-across-restarts behaviour.

— Names the env var and states the before/after default clearly. v26.9.3

THINNER COVERAGE BELOW

11 **KV-quant turbo schemes removed** BREAKING 57

The `--kv-quant turbo2` and `turbo4` schemes are removed; only `4` and `8` are supported going forward.

— Exact removed and remaining values named, migration is implicit. v26.9.3

12 **Concurrent speculative decoding for Flash Next** IMPROVED 56

Flash Next can now draft and verify multiple concurrent chats together, giving +19 to 30% throughput per user at four concurrent chats and +27% at two.

— Concrete throughput numbers but no configuration surface to invoke it. v26.9.3

13 **Apple Intelligence on-device model option** NEW 55

Adds Apple Intelligence as an on-device option in the model picker, shown only when Apple Intelligence is enabled on macOS; supports tools and a 4k context window.

— UI path and requirements named, no deeper mechanism given. v26.9.3

14 **FLUX.2 klein guidance and negative prompt parameters** IMPROVED 55

FLUX.2 klein base now accepts `guidance_scale` and `negative_prompt` parameters.

— Names both parameters but no example or scope detail. v26.9.3

15 **Embedded engine updates and ds4 model routing** IMPROVED 55

Updates embedded engines to llama.cpp v0.4.0 (b10809) and the ds4 September 14 head; ds4's own GGUFs (DeepSeek V4.1 Flash, GLM 5.3 Flash, Qwen3.8 Flash Next) now route to ds4.

— Exact versions and model names but no user-facing action described. v26.9.3

16 **SSM checkpoint cap default reduced** BREAKING 53

`--ssm-checkpoint-max` default changed from 32 to 16, reducing memory use but altering existing default behaviour for models relying on the prior cap.

— Exact before/after values given without further mechanism. v26.9.3

17 **3.3-bit quantized Qwen3.8 Flash Next pack** NEW 50

Adds the `ddalcu/Qwen3.8-Flash-Next-MLX-Serve-iQ-MLX-3.3bpw` 3.3-bit quantized pack for Qwen 3.8 Flash Next, fitting in 52 GB resident on a 64 GB Mac.

— Names the exact pack and memory footprint but not how to load it. v26.9.3

└──▷ **BREAKING ON UPGRADE**

! The `--kv-quant turbo2` and `turbo4` schemes are removed; only `4` and `8` are supported.

! `--ssm-checkpoint-max` now defaults to 16 (previously 32), reducing memory use but changing existing default behaviour.

! The speculative cost table is no longer persisted across restarts unless `MLX_SERVE_ROUND_COST_PERSIST=1` is set.

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

Pinecone

Feedback ⓘ 1 RELEASE · SEEN 2026-09-16 NOTES ↗ RANK

Pinecone is a managed vector database that stores and queries embeddings for AI applications.

> Pinecone introduced a new Documents API with schema-based indexes that replaces the prior records model, alongside a new dedicated read nodes index type with migration and scaling support, a broadened admin API covering organizations, projects, service accounts, API keys and invites, and smaller additions like prefix-based document listing and partial field patching.

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

01 **Dedicated read nodes index type** NEW 80

A new dedicated read nodes index type supports configurable shards and replicas, and can be created from scratch or from a backup of an existing index. Existing on-demand or pod-based indexes can be migrated to this type, and a dedicated read nodes index can be scaled afterward by adding replicas (for query throughput) or shards (for storage capacity), or by adding a hosted embedding model.

Tune a dedicated read nodes index to favour lower latency over perfect recall in a high-throughput production query path.

```
$ curl -X POST 'https://<index-host>/query' \
  -H 'Api-Key: <PINECONE_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"vector": [0.1, 0.2, ...], "topK": 10, "scan_factor": 2,
"max_candidates": 500}'
```

— Names shards/replicas mechanism and includes tunable query example snapshot-20260916

02 **Admin API for orgs, projects, service accounts, keys, invites** NEW 75

Pinecone expanded its administrative API surface across five resource types: service account management endpoints (create, get, delete, update, and rotate OAuth client secret, with the client secret returned only on create or rotate); project management endpoints (list projects, create project, get project details, delete project, and update project name, Pod limit, or CMEK); organization management endpoints (list organizations, get organization details, delete organization, and update organization name); API key management endpoints (list and delete API keys within a project); and an invite management endpoint to list pending and expired invites in an organization.

— Enumerates every admin verb but gives no exact endpoint paths `snapshot-20260916`

03 **Prefix filter and pagination for List Documents** IMPROVED

70

The List Documents endpoint gains a `prefix` parameter to filter results to documents whose IDs start with a given string, with a `pagination` token to page through results.

Page through all documents in a namespace whose IDs share a common prefix — useful for auditing or bulk-exporting a tenant's records.

```
$ curl -X GET 'https://<index-host>/documents?prefix=tenant-acme-
&limit=100' \
  -H 'Api-Key: <PINECONE_API_KEY>'
```

— Named parameters plus a runnable curl example `snapshot-20260916`

THINNER COVERAGE BELOW

04 **Partial field patching on documents** NEW

25

Adds the ability to patch individual fields on existing documents or records without requiring a full replacement.

— One-line description with no mechanism or example given `snapshot-20260916`

└──▷ **BREAKING ON UPGRADE**

! API version `2026-07` introduces the Documents API with schema-based indexes; existing code using the prior records model must be migrated — review Adopt the Documents API to determine if your code is affected.

WAS THIS USEFUL? 👍 👎

okf-memory OKF Agent Memory

Feedback ⓘ

1 RELEASE · 2026-09-16 NOTES ↗ RANK

Git-native persistent memory for AI coding agents. 2 with sub-300µs in-memory BM25 search, embedded MCP server, and progressive disclosure.

> OKF Agent Memory's only release this window ships a new `okf-benchmark` CLI for empirically measuring Dual-Memory Agent Architecture token savings and latency against live LLM endpoints, alongside CLI hardening and the removal of dry-run/mock modes.

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

01 **okf-benchmark DMAA measurement suite** NEW 95

Adds `okf-benchmark` CLI with `-suite push`, `-suite pull`, and `-suite dmaa` subflags to empirically measure token savings and TTFT latency of Dual-Memory Agent Architecture against live LLM endpoints. Supports `-p` (provider) and `-m` (model) flags for OpenAI (`-p openai`) and Anthropic (`-p claude`) endpoints alongside the default local LM Studio server at `http://localhost:1234`, plus `-warmup=true` to prime GPU compute pipelines, Metal context buffers, and KV-caches before measurement for fair prefill latency comparisons.

Run the full end-to-end DMAA benchmark against a local LM Studio instance to measure token savings and TTFT speedup on your own hardware.

```
$ go run ./cmd/okf-benchmark -suite dmaa -warmup=true
```

Benchmark DMAA token and latency impact against a specific cloud model when local GPU hardware is unavailable.

```
$ go run ./cmd/okf-benchmark -p openai -m gpt-4o -suite dmaa -warmup=true
```

Run only the Layer 1 (Working Memory steering tax) benchmark against Anthropic Claude to isolate the AAG vs. prose rules comparison.

```
$ go run ./cmd/okf-benchmark -p claude -m claude-3-7-sonnet-20250219 -suite push
```

— Full flag set, providers, and runnable commands given v0.3.1

02 **Contextual help for okf subcommands** IMPROVED 65

Adds dedicated help handlers for `okf create --help`, `okf update --help`, `okf search --help`, and `okf show --help`, displaying contextual usage instead of creating literal `--help.md` files.

— Exact commands named, minimal further detail v0.3.1

03 **Argument boundary hardening for concept IDs** IMPROVED 60

Adds argument boundary hardening to `okf` subcommands, rejecting hyphenated arguments (e.g. `--invalid-flag`) as concept IDs to prevent accidental concept creation or directory corruption.

— Explains risk prevented, no example command v0.3.1

04 **Removal of dry-run and mock fallbacks** BREAKING 60

Removes synthetic `--dry-run` and mock data fallbacks from `okf-benchmark`; the runner now strictly evaluates live models and exits with clean diagnostics on unreachable endpoints. Any automation or scripts relying on dry-run mode will now fail or exit with diagnostics when endpoints are unreachable.

— Clear before/after behavior, no migration path given v0.3.1

THINNER COVERAGE BELOW

05 **Mermaid diagram AST sanitizer in benchmark output** NEW 55

okf-benchmark includes a built-in Mermaid diagram AST sanitizer that repairs unquoted labels with braces/quotes, normalizes subgraph IDs, and strips empty whitespace lines for error-free Markdown rendering.

— Mechanism described but no invocation shown v0.3.1

└──▷ **BREAKING ON UPGRADE**

! The `--dry-run` flag and mock data fallbacks have been removed from `okf-benchmark`; any automation or scripts that relied on dry-run mode will now fail or exit with diagnostics when endpoints are unreachable.

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ 1 RELEASE · 2026-08-20

NOTES ↗ RANK

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

> LangSmith's biggest changes this window are API-level: a new paginated experiment comparison endpoint replacing legacy helpers, expanded annotation-queue item support for threads, and several breaking permission and endpoint changes, alongside smaller evaluator and dataset-table UI work.

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

01 Bulk dataset split management in experiment tables NEW

50

Select multiple rows (or all matching current filters) in experiment tables to add, replace, or remove dataset splits in one action, or copy selected examples to another dataset. Each split chip in a row's Splits cell in experiment results and comparison views is now interactive, with an Edit splits action opening the single-example split picker.

— UI actions described but only as navigation, no config or endpoint. snapshot-20260916

02 Download datasets permission split from read permission

50

BREAKING

The `download datasets` permission is split from `read datasets`; users previously able to read datasets may no longer be able to download them until the new permission is granted.

— Permission names given but no migration steps or UI path. snapshot-20260916

03 Code evaluator timeout and failure handling IMPROVED

49

Code evaluators now receive more time to run each batch, reducing timeouts for evaluators that import heavy libraries such as scikit-learn. Custom code evaluators that time out or fail on a run now record a visible error on that run instead of silently leaving it without feedback.

Update an existing code evaluator rule in place (no delete-before-create gap) when uploading a revised evaluator.

```
$ langsmith evaluator upload --replace evaluator.py
```

— Behaviour change described, no timing numbers or exact flags. snapshot-20260916

04 Dataset table text truncation and image thumbnails IMPROVED

33

Dataset and experiment tables now truncate long input and reference-output text and show detected base64 images as small thumbnails with a delayed larger preview.

— Describes the display change but no config or path to act on. snapshot-20260916

05 Evaluator spend chart layout on narrow screens IMPROVED

20

Evaluator spend charts on project and dataset evaluator tabs now scroll horizontally on narrow screens instead of compressing the chart and stat cards.

— Bare UI display fix with no further mechanism. snapshot-20260916

└──▷ BREAKING ON UPGRADE

! The legacy feedback formula endpoints (`POST/GET /feedback/formulas` , `GET/PUT/DELETE /feedback/formulas/{feedback_formula_id}`) are scheduled for removal on 2026-08-20; migrate existing feedback formulas to composite evaluators before that date.

! Legacy dataset comparison helpers are removed from the public OpenAPI spec and generated SDKs; use `POST /v2/datasets/<id>/experiment-runs` instead (existing HTTP routes continue to work for LangSmith UI clients).

! The `download datasets` permission is split from `read datasets` ; users previously able to read datasets may no longer be able to download them until the new permission is granted.

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ 1 RELEASE · 2026-09-01 NOTES ↗ RANK

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

> Braintrust's latest bt CLI releases add columnar metric storage for high-volume trace projects, a new built-in GLM-5.3 Flash model, session tracing for Grok and Google Antigravity coding agents, and configurable app public URLs, alongside removal of a long-deprecated CLI flag.

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

01 **GLM-5.3 Flash native inference model** NEW 90

Adds GLM-5.3 Flash as a built-in multimodal reasoning model served via Braintrust's native inference; select it under the Braintrust provider or request `glm-5.3-flash` through the Braintrust Gateway with no AI provider setup required.

— Names exact model id and two ways to invoke it. snapshot-20260916

02 **Configurable app public URL for bt CLI** NEW 80

Adds `--app-public-url` flag and `BRAINTRUST_APP_PUBLIC_URL` environment variable (bt v0.19.3+) for deployments where the browser-facing app URL differs from the authentication endpoint.

— Exact flag and env var name given, directly usable. snapshot-20260916

03 **Columnar metric storage for traces** NEW 75

Adds a Metric columnstore, enabled in project settings, which stores trace-level metrics in a columnar format to reduce data read on high-volume projects; requires data plane v2.2.1+.

— Names mechanism, config location, and version requirement. snapshot-20260916

04 **Removal of --relabel-overlap flag** BREAKING 75

The `--relabel-overlap` flag is removed from `bt topics config enable` and `bt topics config set` in bt v0.20.0; the flag was deprecated and ignored since data plane v2.15.0.

— Names exact commands and versions but no migration step. snapshot-20260916

THINNER COVERAGE BELOW

05 **Session tracing for Grok and Google Antigravity** NEW 55

The `bt` CLI (v0.19.3+) adds Grok session tracing, capturing each coding session as a Braintrust trace without interrupting the agent if tracing fails, and adds equivalent

session tracing for Google Antigravity sessions.

— Describes behaviour but not setup steps or config. snapshot-20260916

└──▷ **BREAKING ON UPGRADE**

! The `--relabel-overlap` flag is removed from `bt topics config enable` and `bt topics config set` in bt v0.20.0; the flag was deprecated and ignored since data plane v2.15.0.

WAS THIS USEFUL?  

ai-gateway

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗ RANK

AI Gateway routes AI model requests through a self-hosted gateway with observability and security controls.

> ai-gateway v1.5.7 introduces a set of content guardrail plugins — pii-redact, schema-guard, secret-scan, regex-guard, and prompt-shield — built on new shared plugin APIs, alongside stricter startup-time validation that now rejects previously-silent misconfigurations.

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

01 Stricter startup validation for guardrail plugin configs

91

BREAKING

`ferrogw validate` and `ferrogw doctor` now report all startup-time errors for content guardrails — misspelled `action`, uncompilable patterns, unrecognised entities/kinds/categories, and malformed rule or schema keywords — before deploy rather than at a failed start. A plugin listed at a stage it cannot act at now fails the load instead of registering silently and returning early on every request (`pii-redact` and `prompt-shield` require `before_request`; `schema-guard` requires `after_request`); an `action` set to an unrecognised value, a `rules` block that is not a list, an unrecognised `apply_to`, an empty selector (`rules: []`, `entities: []`, `kinds: []`, `categories: []`), or a key written with the wrong value kind (e.g. `action: 1`) are now startup errors; and an empty entry inside a `patterns` list on `pii-redact` or `secret-scan` is now a startup error, since a blank pattern previously matched every string and caused all requests to be denied or rewritten.

— Names exact commands and every config condition that now fails. v1.5.7

02 pii-redact PII redaction plugin

84

BREAKING

With `action: "redact"`, `pii-redact` rewrites requests in place on `/v1/chat/completions` (streamed or not) and `/v1/completions`; detections on other surfaces (`/v1/embeddings`, `/v1/images/generations`, `/v1/responses`, `/v1/rerank`, `/v1/moderations`, `/v1/audio/*`, `/v1/*` pass-through) are denied instead. Its `redact_placeholder` is now inserted as literal text instead of a replacement template — a placeholder that previously contained `$0` (which wrote the detected value back through) or any dollar sign is now forwarded verbatim to the provider.

— Endpoints and config key named, with explicit before/after behaviour. v1.5.7

03 Shared plugin APIs for text scanning and validation NEW

84

Adds `plugin.RequestText` / `plugin.ResponseText` iterators that walk every text-bearing surface of a request or response — content parts of every type, reasoning content, and each tool call's arguments — in both directions (`Part.ImageURL` and function names are deliberately excluded); `plugin.RejectUninspectable` as the shared verdict for a `before_request` whose content could not be projected as text; `plugin.NormalizeAction` and `plugin.ValidateAction` to canonicalise and validate a configured `action` against the closed set a plugin can honour (values carrying a `${VAR}` reference are passed rather than judged); `plugin.StageRestricted` / `plugin.ValidateStage` so a plugin can declare the stages it acts at, with `Manager.Register` refusing any undeclared stage; and `plugin.StringSetting` / `plugin.ListSetting` to read an optional config key, distinguishing an absent key from one written with the wrong value kind.

— Every API named with its exact behaviour, aimed at plugin authors. v1.5.7

---

04 **schema-guard JSON Schema validation plugin** NEW 77

`schema-guard` validates one assembled document per choice and runs at `after_request` only; it handles `n > 1` responses, rejects empty choices, and passes tool-call-only choices through without schema validation.

Validate that LLM responses conform to a JSON Schema before returning them to the client, catching empty or structurally invalid answers.

yaml

```yaml
plugins:
  - name: schema-guard
    stage: after_request
    config:
      action: block
      schema:
        type: object
        required:
          - answer
        properties:
          answer:
            type: string
          confidence:
            type: number
```

— Behaviour and edge cases named, with a runnable example config. v1.5.7

THINNER COVERAGE BELOW

05 **secret-scan detection pattern coverage** IMPROVED 58

`secret-scan`'s `github_token` kind now matches fine-grained personal access tokens (`github_pat_…`) in addition to classic prefixes. `slack_token` also matches

app-level (`xapp-…`) and rotation (`xoxe-…`) tokens; `private_key` also matches DSA, ENCRYPTED, and OpenPGP armor (`PRIVATE KEY BLOCK`) PEM headers; `stripe_key` also matches webhook signing secrets (`whsec_…`).

— Names every added pattern but gives no config surface to act on. v1.5.7

---

06 **regex-guard and prompt-shield guardrail plugins** NEW

50

Adds `regex-guard` and `prompt-shield` as new registered content guardrail plugins, importable via blank import. `regex-guard` 's `rules` key is required and must name at least one rule — omitting it or providing an empty list is a startup error — and an entry whose rules all carry `apply_to: output` fails at `before_request` ; `prompt-shield` requires the `before_request` stage.

— Named and given basic constraints but little functional description. v1.5.7

└──▷ **BREAKING ON UPGRADE**

! A plugin listed at a stage it cannot act at now fails the load (and is reported by `ferrogw validate`) instead of registering silently and returning early on every request — `pii-redact` and `prompt-shield` require `before_request` ; `schema-guard` requires `after_request` ; a `regex-guard` entry whose rules all carry `apply_to: output` fails at `before_request` .

! `pii-redact` 's `redact_placeholder` is now inserted as literal text instead of a replacement template — a placeholder that previously contained `$0` (which wrote the detected value back through) or any dollar sign will now be forwarded verbatim to the provider.

! An `action` set to an unrecognised value, a `rules` block that is not a list, a `regex-guard` `apply_to` that is unrecognised, an empty selector (`rules: []` , `entities: []` , `kinds: []` , `categories: []`), or a key written with the wrong value kind (e.g. `action: 1`) are all startup errors that prevent the plugin from loading.

! `regex-guard` 's `rules` key is required and must name at least one rule — omitting it or providing an empty list is a startup error.

! An empty entry inside a `patterns` list on `pii-redact` or `secret-scan` is now a startup error (a blank pattern previously matched every string, causing all requests to be denied or rewritten under `block` / `redact`).

WAS THIS USEFUL? 👍 👎

Langfuse

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗ RANK

Langfuse provides tracing, evaluation, and monitoring for LLM applications.

> Langfuse v4.37.0 adds a bulk dataset ingestion tool for MCP, expands AI gateway observability with generation uploads and operational telemetry, and raises the blob storage object size cap.

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

01 **batchUpsertDatasetItems MCP tool** NEW 55

The MCP server gains a `batchUpsertDatasetItems` tool for bulk dataset item ingestion.

— Names exact tool but no usage detail or limits v4.37.0

02 **S3 multipart part size for blob storage** IMPROVED 50

Raises the S3 multipart part size to lift the 48.8 GiB object cap for blob storage.

— Gives concrete before limit but no config key or new value v4.37.0

03 **AI gateway observability and generation uploads** NEW 45

The AI gateway now uploads finalized generations to Langfuse for tracing and audit, gains operational logs, traces, and metrics for gateway-level observability, and captures Responses API facts in debug logs.

— Describes what changed but no config keys or endpoints given v4.37.0

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI/LLM Security

ai-safe2-framework

Feedback ⓘ 1 RELEASE · 2026-09-15

NOTES ↗ RANK

Ai-Safe2-Framework provides governance, risk, and compliance controls for securing agentic AI systems and non-human identities.

> AI SAFE² CLI 0.5.0 introduces an evidence-driven failure diagnosis pipeline for agentic systems, alongside scanner hygiene and a new agent delivery workflow guide.

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

01 **Evidence diagnose command for agent failure localization** NEW

95

Adds `safe2 evidence diagnose` to rank evidence-supported failure locations across an identified agent system, with `--output`, `--card <file>.md`, and `--strict` flags. The command consumes the new `safe2.failure-source.v1` schema (a strict, provider-neutral contract for observations and competing failure candidates) and produces the `safe2.failure-diagnosis.v1` schema, returning deterministic candidate rankings, evidence assessments, ownership assignments, and recommended actions. `--strict` returns a nonzero exit code when evidence cannot support one strong or moderate leading candidate, and `--card` emits a human-readable Failure Localization Card in Markdown alongside the JSON diagnosis. A companion `safe2 schema validate failure-diagnosis-v1` subcommand validates a produced diagnosis contract against the versioned schema.

— Names exact command, flags, schemas, and validation subcommand

2026-09-15_CLI_0.5.0_Find_Where_Agent_W…

THINNER COVERAGE BELOW

02 **Scanner exclusions for generated evidence directories**

IMPROVED

45

Adds scanner workspace exclusions to avoid traversing generated AI SAFE² evidence directories, named virtual environments, and uv caches during repository scans.

— Describes behavior change but no config key or flag given

2026-09-15_CLI_0.5.0_Find_Where_Agent_W…

03 **Portable agent delivery workflow guidance** NEW

30

Adds a portable agent delivery workflow to guide agents through isolation, building, proof, self-assessment, independent review, and shipping with receipts.

— Lists workflow stages but no concrete command or config surface

2026-09-15_CLI_0.5.0_Find_Where_Agent_W…

WAS THIS USEFUL? 👍 👎

FailproofAI

Feedback ⓘ 1 RELEASE · 2026-09-16 NOTES ↗ RANK

Observability and enforcement for AI agent harnesses. Capture every run and runtime reliability with policy enforcement.

> FailproofAI 1.0.6 brings native pre-tool-call policy enforcement to Hermes on par with its existing OpenClaw integration, extending coverage to all active Hermes profiles with daemon readiness checks, canonical evaluation metadata, and a socket deadline for stalled connections.

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

01 Multi-profile install and health checks for Hermes IMPROVED

75

Extends installation, removal, and health checks to all active Hermes profiles: the default `~/.hermes` profile, upstream nested profiles, and configured `~/.hermes-<name>` profiles. Running `npm install -g failproofai@1.0.6 && failproofai config` after upgrade discovers all active profiles, verifies daemon readiness, and installs the native policy plugin into each.

After upgrading, run `failproofai config` so FailproofAI discovers all active Hermes profiles, verifies daemon readiness, and installs the native policy plugin into each one.

```
$ npm install -g failproofai@1.0.6 && failproofai config
```

— Exact profile paths and a runnable upgrade command given 1.0.6

02 Native pre-tool-call policy enforcement for Hermes NEW

70

Adds native, pre-tool-call policy enforcement for Hermes via a profile-local plugin backed by the `failproofaid` warm worker, bringing Hermes in line with the existing OpenClaw integration. Supports deny() as a deterministic hard block and instruct() as a model-visible advisory, with a bounded retry ledger to prevent advisory loops from deadlocking a turn.

— Mechanism and functions named but no direct enable command shown 1.0.6

THINNER COVERAGE BELOW

03 Daemon readiness check before Hermes config changes NEW

55

Adds a daemon readiness check that verifies the local daemon supports the exact native policy protocol before modifying Hermes configuration, halting installation with setup

guidance when the daemon is missing, unhealthy, outdated, or protocol-incompatible.

— Behavior detailed but no specific command or flag named `1.0.6`

04 **Canonical metadata fields for Hermes policy evaluations** NEW 40

Policy evaluations for Hermes now emit canonical metadata including `tool`, `policy`, `match`, `reason`, and `latency` fields.

— Fields named but no way to access or consume them given `1.0.6`

05 **Socket deadline for Hermes policy evaluation** NEW 35

Applies a single total socket deadline to Hermes policy evaluation to limit the effect of stalled or partially responsive daemon connections.

— Behavior described briefly with no configurable value shown `1.0.6`

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Anthropic

Feedback ⓘ 1 RELEASE · SEEN 2026-09-16 NOTES ↗ RANK

Anthropic provides Claude AI models and APIs for building applications that generate, analyze, and automate work.

> Anthropic's Messages API added on-demand conversation compaction, an `auto` permission policy for Managed Agents tool calls, a new CLI command to attach to live agent sessions, and extended per-message effort controls to Google Cloud for additional models.

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

01 On-demand conversation compaction in Messages API NEW

90

Adds on-demand conversation compaction to the Messages API (beta) via the `compact-2026-09-04` beta header and a top-level `compaction` parameter. The API returns a signed compaction block that replaces earlier messages on subsequent requests, letting you control when to compact, keep recent turns verbatim, and preserve thinking in kept turns on supported models.

— Names beta header and parameter but no full request example. snapshot-20260916

02 Auto permission policy for Managed Agents tool calls NEW

80

Adds `auto` as a permission policy option in Claude Managed Agents, where the server evaluates each agent or MCP tool call and runs, denies, or pauses it for approval. Evaluation outcomes are reported in the `evaluation` field alongside `evaluated_permission` on `agent.tool_use` and `agent.mcp_tool_use` events.

— Names fields and events but not full setup steps. snapshot-20260916

03 CLI session attachment for Managed Agents NEW

80

Adds `ant beta:sessions connect` to the `ant` CLI, attaching a terminal to a running Claude Managed Agents session to follow it live, send messages, and allow or deny pending tool calls. Pass `--web` to serve the Claude Console session viewer locally instead.

— Gives a runnable CLI command and flag. snapshot-20260916

04 Per-message effort changes on Google Cloud IMPROVED

65

Extends per-message effort changes (beta) to Google Cloud for Claude Fable 5.1, Claude Mythos 5.1, and Claude Opus 5, using the existing `mid-conversation-output-`

`config-2026-07-01` beta header.

— Names header and models but no usage example. snapshot-20260916

WAS THIS USEFUL? 👍 👎

Google Gemini API

Feedback ⓘ 1 RELEASE · 2026-09-15 NOTES ↗

RANK

Gemini API provides Gemini models for generating text, images, and other multimodal content in applications.

> Gemini API launched general availability of two new Live API audio-to-audio models built for real-time voice agents, one optimized for low latency and one for extended background reasoning.

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

01 **gemini-3.8-live model in Live API** NEW 78

The `gemini-3.8-live` model is now generally available via the Live API: a low-latency audio-to-audio model for real-time voice agents, featuring interleaved reasoning, asynchronous function calling enabled by default, and full session client content updates.

— Names model id and lists concrete mechanisms; no usage snippet given September 15, 2026

THINNER COVERAGE BELOW

02 **gemini-3.8-live-extended-thinking model in Live API** NEW 53

The `gemini-3.8-live-extended-thinking` model is now generally available via the Live API: a high-reasoning audio-to-audio model that supports background reasoning during live audio interactions.

— Names model but describes mechanism only briefly September 15, 2026

WAS THIS USEFUL? 👍 👎

◆ Open-Weight Models

Black Forest Labs FLUX

Feedback ⓘ 1 RELEASE · 2026-09-10

NOTES ↗ RANK

FLUX is Black Forest Labs' family of visual generation and editing models.

> FLUX 3 Video now supports 2K and 4K output directly from the API, removing the need for a separate upscale step.

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

01 **2K/4K resolution options for FLUX 3 Video** NEW 83

Adds `qhd` (2560×1440) and `uhd` (3840×2176) values to the `resolution` field on `POST /v1/flux-3-video`, enabling 2K and 4K video output in a single request without a separate upscale call. Support for `qhd` and `uhd` also extends to `draft_enhance`, allowing drafts to be committed directly to 2K or 4K, though drafts themselves still render at `hd`.

— Names exact endpoint, field, values and resolutions but no examples.

FLUX 3 Video — 2K and 4K Output

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · September 16, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue