THE AI TOOLCHAIN

NO. 025

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED SEPTEMBER 13, 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 139 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 21 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. Braintrust's Patterns mines your traces for recurring failure modes and turns them into scorers, so you reading runs by hand to find what to test. OpenRouter gives any

vim ? help

model a hosted sandboxed Linux shell without you running the container. And Anthropic puts approval gates on managed agents before they act.

> EVALUATE
>
> **Find the recurring failure in your traces without reading them yourself**
>
> Braintrust's Patterns clusters recurring problems across logged traces and lets you promote a finding straight into a scorer or prompt, closing the usual gap where someone eyeballs bad runs and hand-writes an eval. LangSmith's public `POST /v2/datasets/<id>/experiment-runs` puts experiment comparison behind an API, so a regression check between two prompt or model versions can run in CI instead of in the UI.
>
> Braintrust · LangSmith

> GOVERN
>
> **Require a human approval before an agent takes a consequential action**
>
> Claude Managed Agents now carry human-in-the-loop permission policies plus declarative infra-as-code apply and live session control, so agent authority is defined in a reviewed file rather than in whoever launched the run. OpenAI's API key expiration controls mean a credential handed to an agent or contractor stops working on its own instead of living forever in an env var.
>
> Anthropic · OpenAI

> BUILD
>
> **Give a model a real shell without standing up your own sandbox**
>
> OpenRouter's hosted shell server tool runs isolated Linux commands and returns output to any model on the Responses and Messages APIs, removing the container-per-session plumbing teams currently build to let agents actually execute things. Fireworks' training skill lets a coding agent plan and price a fine-tuning job before it launches, so cost is known up front rather than on the invoice.
>
> OpenRouter · Fireworks AI

OFFENSIVE SECURITY

◆ Exploitation & C2

xalgorix

Feedback ⓘ 2 RELEASES · 2026-09-12 → 2026-09-13 NOTES ↗

RANK

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

> Xalgorix shipped strict proxy enforcement with credential masking, a dual-agent exploit verification architecture that re-confirms every finding before it reaches a report, Telegram alerting, expanded multi-arch Docker and native macOS binaries, a portable wordlist fallback for common fuzzing tools, and hardened report logo handling.

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

01 Strict proxy enforcement via `XALGORIX_PROXY_REQUIRED` NEW

88

Adds `XALGORIX_PROXY_REQUIRED=true` to enforce that all scan traffic routes strictly through a single upstream proxy, failing immediately on startup if `XALGORIX_USE_PROXY=true` or `XALGORIX_PROXY_URL` are missing or invalid. Runs a local loopback HTTP bridge for Chromium headless instances and proxy-aware subprocesses (`terminal_execute` , `python_action`) to eliminate proxy credential exposure in CLI process arguments, masks upstream proxy credentials across logs, errors, and process environments, and registers the new variable in the settings environment schema at `internal/web/settings_env.go` with guidance in `docs/proxy-support.md` .

— exact env var, file paths and failure behaviour all named v4.6.76

02 Portable wordlist resolution for fuzzing tools NEW

88

Materializes a fallback embedded wordlist at `tmp/wordlists/xalgorix-web-common.txt` so prebuilt installs can run `ffuf` , `gobuster` , `feroxbuster` , `dirsearch` , `dirb` , and `wfuzz` without a separate SecLists package, while preserving existing, generated, relative, and variable-based wordlists as well as ffuf custom keyword suffixes, and disclosing any substitution in the tool result.

— exact fallback path and full tool list with fallback behaviour v4.6.75

03 Telegram bot alerts and scan-complete notifications NEW

86

Adds `XALGORIX_TELEGRAM_BOT_TOKEN` , `XALGORIX_TELEGRAM_CHAT_ID` , and `XALGORIX_TELEGRAM_MIN_SEVERITY` environment variables to enable Telegram bot

notifications with severity-gated real-time finding alerts and automatic PDF report attachment delivery, plus `XALGORIX_NOTIFY_SCAN_COMPLETE=true` to deliver executive summaries to Discord and Telegram when a scan finishes.

Route high-severity Telegram alerts to a private channel and get a PDF summary when a scan completes — useful for unattended overnight engagements.

```text
XALGORIX_TELEGRAM_BOT_TOKEN=123456789:ABCdefGhIJKlmNoPQRsTUVwxyZ
XALGORIX_TELEGRAM_CHAT_ID=-1001234567890
XALGORIX_TELEGRAM_MIN_SEVERITY=high
XALGORIX_NOTIFY_SCAN_COMPLETE=true
```

Route high-severity findings and the final scan report directly to a Telegram channel so the team gets confirmed, verified exploits without polling the dashboard.


```text
XALGORIX_TELEGRAM_BOT_TOKEN=123456789:ABCdefGhIJKlmNoPQRsTUVwxyZ
XALGORIX_TELEGRAM_CHAT_ID=-1001234567890
XALGORIX_TELEGRAM_MIN_SEVERITY=high
XALGORIX_NOTIFY_SCAN_COMPLETE=true
```


— exact env vars with a runnable config example `product docs`

04 **Expanded platform and architecture support** NEW 75

Publishes official multi-architecture Docker images for Linux `amd64` and `arm64` via `ghcr.io/xalgord/xalgorix:latest` with `docker-compose` support, and adds native macOS ARM64 (Apple Silicon) binary releases alongside existing Linux and WSL-on-Windows support.

Deploy Xalgorix on an ARM64 server (e.g. Graviton or Apple Silicon) using the official multi-arch image with your env file mounted.

```
$ docker run --rm -it --network host -v
~/.xalgorix.env:/root/.xalgorix.env
ghcr.io/xalgord/xalgorix:latest
```

Run the containerised engine on an arm64 server (e.g. AWS Graviton) without building from source.

```
$ docker run --rm -it --network host -v
~/.xalgorix.env:/root/.xalgorix.env
ghcr.io/xalgord/xalgorix:latest
```

— runnable docker command and named platforms given `product docs`

05 **Hardened report logo loading** IMPROVED 65

Confines report logos to validated image data loaded from the configured uploads directory, enforcing supported image formats, a 5 MiB read limit, symlink escape rejection, and prevention of PDF renderers reopening user-supplied paths.

— concrete limits and protections named but no user-facing action `v4.6.75`

06 **Dual-agent exploit verification architecture** NEW 60

An autonomous agent executes a 22-phase penetration testing methodology while an independent verifier agent re-exploits every finding before it is included in the scan record or PDF report, rejecting non-reproducible results.

— mechanism explained but no config or API surface to act on `product docs`

THINNER COVERAGE BELOW

07 **Dashboard model selector and theme persistence** NEW 50

Adds a unified provider model selector — a reusable model field across API keys, OAuth, and local endpoints — in the dashboard, and adds Light, Dark, and System dashboard theme persistence with no-flash startup.

— names the UI surfaces but no config keys or commands `product docs`

08 **MiniMax native web search integration** NEW 25

Adds MiniMax native web search integration for MiniMax-backed scans.

— one-line mention with no mechanism or scope `product docs`

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

Warp

Feedback ⓘ 1 RELEASE · 2026-09-13 CHANGELOG ↗ RANK

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

> Warp's latest changelog introduces Grok Build support, fzf.fish history and file handoff, and team-aware Oz CLI commands, alongside refinements to the WASM conversation viewer and team switching on Warp on Web.

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

01 **In-window team switching on Warp on Web** IMPROVED 40

On Warp on Web, switching teams now updates the current window instead of replacing the session.

— Clear before/after but no further scope changelog-20260913-9cd42abb

02 **Web top bar in WASM conversation viewer** IMPROVED 35

The WASM conversation viewer now shows the web top bar for orchestrated factory sessions.

— Names the surface but no mechanism detail changelog-20260913-9cd42abb

03 **Grok Build support** NEW 15

Warp adds Grok Build support.

— Named only in summary, no further detail changelog-20260913-9cd42abb

04 **fzf.fish history and file handoff** NEW 15

Warp adds fzf.fish history/file handoff.

— Named only in summary, no further detail changelog-20260913-9cd42abb

05 **Team-aware Oz CLI commands** NEW 15

Warp adds team-aware Oz CLI commands.

— Named only in summary, no further detail changelog-20260913-9cd42abb

WAS THIS USEFUL? 👍 👎

StackBlitz bolt.new

Feedback ⓘ 1 RELEASE · 2026-09-13

CHANGELOG ↗ RANK

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

> Bolt.new added direct visual editing in the live preview, letting users tweak text and styling without writing prompts.

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

01 **Direct visual editing in preview** NEW 60

Users can edit text, colors, font size, formatting, and alignment directly on the preview canvas without prompting the AI. Changes are batched above the chatbox and only written to the project code once `Save changes` is clicked.

Iterate on a project's look without spending tokens — make visual edits in the preview, then commit them all at once.

> 📍 Click the visual-edit tool in the chatbox, adjust text, colors, and font styles directly on the preview canvas, then click `Save changes` when satisfied to write the updates to your project code.

— Names UI flow and save mechanism but no config surface. changelog-20260913-050f2bb3

WAS THIS USEFUL? 👍 👎

Vercel v0 Feedback ⓘ changelog · Sep 13, 2026 CHANGELOG ↗

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

v0: Adds a Default Chat Visibility setting under Settings → Workspace to control whether new team chats default to Private, Team can view, or Team can edit.

└──▷ **HOW TO FIND IT**

Lock down a workspace so all new chats start private — useful for teams that share sensitive work and want explicit sharing opt-in.

> 📍 In the v0 console, go to Settings → Workspace → Default Chat Visibility and select Private.

Open collaboration by default so every new chat is immediately editable by the whole team, removing the need for per-chat invite steps.

> 📍 In the v0 console, go to Settings → Workspace → Default Chat Visibility and select Team can edit.

New chats in a team workspace are now visible to the team by default, replacing the previous invite-only discovery model.

WAS THIS USEFUL? 👍 👎

Cognition Devin Desktop

Feedback ⓘ 1 RELEASE · 2026-09-13

CHANGELOG ↗ RANK

Devin Desktop is an agent-native code editor for developing software locally, remotely, and in containerized environments.

> Devin Desktop adds SSH-based remote agent hosts, expands Devin Review and PR card actions, and streamlines cloud session and inbox workflows, while removing several legacy Cascade entry points and the always-on ACP toggle.

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

01 **SSH remote agent hosts** NEW 80

Adds Remote agent hosts : connect to agents over SSH from the location selector, with saved hosts, status dots, a server icon in the sidebar, and automatic transcript recovery after reconnect.

Connect to a remote agent running on a cloud VM or jump host without leaving Devin Desktop.

📍 In the location selector, click the location picker › choose Remote agent hosts › click Add host › enter your SSH host details › select the saved host to start an SSH-backed agent session.

— Mechanism and steps given, no config file or flag named changelog-20260913-6da7d82c

02 **Devin Review and PR card enhancements** IMPROVED 80

PR cards gain Open Devin Review , Open in Desktop tab , Close PR , and Git settings actions, and Checkout branch now finds local clones including worktrees. Devin Review adds Cmd/Ctrl+F find-in-page, shows the PR title on review tabs, opens GitHub links at their diff anchor, and swaps the sessions sidebar to a PR files/review sidebar or wiki page tree during fullscreen review.

— Names every PR card action and review surface added changelog-20260913-6da7d82c

03 **Batch inbox actions and navigation** NEW 60

Adds per-row checkboxes in the agent inbox list for batch archive/unarchive, and back/forward arrows in the title bar to navigate between the inbox and opened sessions.

Bulk-archive completed sessions in the inbox after a sprint to keep the workspace clean.

> 📍 In the agent sidebar, open the inbox list view › tick the per-row checkboxes next to the sessions you want to retire › click `Archive` to archive them all at once.

— Names UI controls with a concrete walkthrough changelog-20260913-6da7d82c

THINNER COVERAGE BELOW

04 **Legacy Cascade entry points removed** BREAKING 55

Several legacy Cascade entry points are removed: new-agent option, annotations, conversation mentions, settings group, automatic Lifeguard checks, Vibe & Replace, Deploy, and workflow commands.

— Lists all removed entry points, no replacement path given changelog-20260913-6da7d82c

05 **File explorer syncs with active session directory** IMPROVED 45

File explorer updates to match the working directory of the active session when switching between sessions in the Agent Command Center.

— Describes behavior and location, no config surface changelog-20260913-6da7d82c

06 **ACP toggle removed, always enabled** BREAKING 45

The `Enable ACP` toggle is removed — ACP is now always enabled and cannot be disabled.

— Names the removed toggle but no migration guidance changelog-20260913-6da7d82c

07 **Cloud session formatting improvements** IMPROVED 40

Cloud sessions gain an image preview carousel, hook run visibility, and rich queued messages in their formatting.

— Names three additions but no mechanism detail changelog-20260913-6da7d82c

08 **Problems panel routes to Devin Local** IMPROVED 40

Problems panel actions send problems to Devin Local when Cascade is disabled.

— Clear behavior stated but minimal mechanism changelog-20260913-6da7d82c

└──▷ ALSO FROM THESE RELEASES

Review a PR diff in-app and update the allowlist when Devin requests network access to unfamiliar domains.

📍 When the `Allowlist N domains` card appears in the session, click it to see the named security profile blocking the destinations › click `Update security profile` to permit the required domains.

└─▷ **BREAKING ON UPGRADE**

! The `Enable ACP` toggle is removed — ACP is now always enabled and cannot be disabled.

! Several legacy Cascade entry points are removed: new-agent option, annotations, conversation mentions, settings group, automatic Lifeguard checks, Vibe & Replace, Deploy, and workflow commands.

WAS THIS USEFUL?  

Letta Code

Feedback ⓘ 2 RELEASES · 2026-09-13 NOTES ↗ RANK

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

> Letta Code's latest releases focus on developer visibility and multi-agent coordination, adding a diagnostic slash command, an agent-to-agent messaging tool, non-blocking Cloud message queuing, and clearer message receipts in the TUI.

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

01 **Non-blocking Cloud message queuing** NEW 75

Adds a `--no-wait` flag to the CLI for enqueuing Cloud messages without blocking until the agent responds, useful for headless/CI invocations like `letta -p --agent <agent-id> --no-wait "Run the nightly report"`.

Fire a headless message to a Cloud-hosted agent from CI without waiting for the response to complete.

```
$ letta -p --agent <agent-id> --no-wait "Run the nightly report"
```

— Names exact flag and shows runnable CLI example v0.32.4

02 **/doctor diagnostic slash command** NEW 65

Adds a `/doctor [symptom]` slash command to investigate agent behavior by analyzing conversation history for clues, e.g. `/doctor agent keeps ignoring my formatting instructions`.

Diagnose why an agent is behaving unexpectedly by letting `/doctor` analyze the conversation history for clues.

```
$ /doctor agent keeps ignoring my formatting instructions
```

— Names command syntax with runnable example v0.32.4

THINNER COVERAGE BELOW

03 **SendAgentMessage for agent coordination** NEW 45

Adds a `SendAgentMessage` tool for direct agent-to-agent coordination in multi-agent workflows.

— Names the tool but no usage example given v0.32.4

04 **Readable message receipts in TUI** IMPROVED 35

Displays readable agent message receipts in the TUI, making it easier to track message delivery status during conversations.

— Describes UI change but no navigation path or mechanism v0.32.6

WAS THIS USEFUL? 👍 👎

Sourcegraph Amp

Feedback ⓘ 1 RELEASE · 2026-09-13 BLOG ↗

RANK

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

> Amp removed its BYOK token fees and rolled out a free Hobby tier and free Teams tier, while also opening up BYOK support to a wide range of external model providers and extending SSO and data-retention protections beyond Enterprise.

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

01 **BYOK support for OpenRouter, Bedrock, Vertex AI, Azure and more**

NEW 85

Adds early-access BYOK support for `OpenRouter`, Amazon Bedrock, Google Cloud Agent Platform (Vertex AI), Azure Foundry, Vercel AI Gateway, Cloudflare AI Gateway, Ollama Cloud, `OpenCode Go`, and custom endpoint URLs — enabled via the More AI Routers & Subscriptions experimental feature and configured under Model Routing in personal or workspace settings.

— Names every provider plus the exact feature flag and settings location.

Free Agent (2026-09-13)

THINNER COVERAGE BELOW

02 **Free SAML/OIDC SSO for workspaces** NEW 50

Grants free SAML/OIDC SSO to the entire workspace when at least one member is on a paid tier.

— Names the exact auth protocols and eligibility condition. Free Agent (2026-09-13)

03 **Free Teams tier with shared collaboration** NEW 45

A new Teams tier is available at no extra charge, enabling free workspace membership for all invitees with shared threads, shared portals, and multiplayer collaboration.

— Names features included but no usage path given. Free Agent (2026-09-13)

04 **BYOK token fees and limits removed** IMPROVED 45

Removes BYOK token fees and limits for all non-Enterprise tiers, allowing use of ChatGPT subscriptions, external providers, or a company AI gateway without per-token charges.

— Clear change in billing but no config detail. Free Agent (2026-09-13)

05 **Free Hobby tier for BYOK users** NEW 42

Amp introduces a free Hobby tier requiring no monthly plan — users can run Amp agents on their own computers at no cost by bringing their own model subscriptions or API keys.

— Describes the tier but no setup steps or limits. Free Agent (2026-09-13)

06 **Zero/minimal data retention extended workspace-wide** IMPROVED 30

Extends zero and minimal data retention policy contractually to all Amp users and workspaces, not just Enterprise.

— States the policy change but no mechanism or scope detail. Free Agent (2026-09-13)

WAS THIS USEFUL?  

Graphify Labs Graphify

Feedback ⓘ 1 RELEASE · 2026-09-12

NOTES ↗ RANK

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

> Graphify's latest releases focus on parser depth and Python 3.14 compatibility, adding new SQL and Rust node types, a symbol disambiguation syntax, and drop-in dependency swaps for the new interpreter.

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

01 **path::Symbol disambiguator in graphify explain** IMPROVED

75

`graphify explain` now accepts a `path::Symbol` form to disambiguate a symbol that shares its name with its file, giving the ambiguity hint a form the resolver actually accepts, e.g. `graphify explain src/auth/session.py::session`.

Disambiguate a symbol whose name collides with its containing file when multiple matches are reported.

```
$ graphify explain src/auth/session.py::session
```

— Runnable command shown but limited to one use case product docs

02 **Python 3.14 support with dependency shims** NEW

60

Graphify now supports Python 3.14, gating optional dependencies that lack 3.14 wheels to drop-in replacements by interpreter version: `jieba` is swapped for `jieba-py` and `graspologic` for `graspologic-native`.

— Names exact package swaps but no install steps shown v0.9.60

THINNER COVERAGE BELOW

03 **New node types: SQL indexes and Rust consts** NEW

25

Graphify adds SQL index node extraction and Rust const nodes to its parsers, expanding the set of code and schema constructs represented in the knowledge graph.

— Named but undescribed beyond the two node types product docs

WAS THIS USEFUL? 👍 👎

Amazon Kiro

Feedback ⓘ 3 RELEASES · 2026-09-05 → 2026-09-11 BLOG ↗

RANK

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

> Kiro's CLI V3 line added a session dashboard, a unified `/config` panel, cloud configuration sync, and settings-navigation cleanups, while Crew 0.6 introduced remote crew chat, selectable agent harnesses, sub-agent dispatching, an Apps Launchpad, and longer unattended run limits alongside several breaking changes including a Python 3.12 minimum.

└──▷ WHAT SHIPPED · 30 FEATURES most completely described first ? what's the number?

01 Configurable timeouts for chat turns and subagents NEW

85

Adds `agent.chat_turn_timeout_secs` (default `14400`) to allow chat turns to run up to four hours, `agent.subagent_timeout_secs` (default `10800`) for three-hour subagents, and `agent.subagent_max_turns` supporting up to 1000 tool calls.

Set per-session and subagent timeouts high enough for multi-hour autonomous work, and cap subagent tool calls to avoid runaway loops.

```json
{
  "agent": {
    "chat_turn_timeout_secs": 14400,
    "subagent_timeout_secs": 10800,
    "subagent_max_turns": 500
  },
  "orchestrator": {
    "max_plan_duration_seconds": 0
  }
}
```

— three keys with exact defaults and limits plus example config

Crew: Remote Crews, Agent Harnesses, an…

02 V3 agent harness polish: Powers picker, per-agent models, MCP refresh, shell approval, stream-json errors

NEW

75

V3 lists installed Powers in the `/` command picker; supports `agent model:` setting so per-agent model preferences apply to models available in the session; updates

available MCP tools when MCP is disabled during an active session; applies shell approval rules to commands run through `logbash` and `logzsh`; and reports an error before exiting when a non-interactive `--output-format stream-json` run is interrupted.

— five concrete surfaces named, little mechanism per item

CLI 2.21.4: Session Dashboard, Configur…

03 **Unified keyboard navigation across settings menus** IMPROVED

75

Unifies keyboard navigation across all settings menus (Display, Status line, Theme, `/verbosity`): Up/Down move between rows, Right/Enter changes a value or opens a submenu, Left changes a value in the other direction, and Esc goes back.

— detailed key mapping across named menus, no command surface

CLI: Session Search Scope, V2 Harness F…

04 **App UI development restricted to install root** BREAKING

75

Developing an app UI outside its install root from the dashboard or API is no longer supported; use `kirocrew app dev <name> --confirm-out-of-install-root` in the terminal instead to develop an app UI outside its install root.

— breaking change paired with exact replacement command

Crew: Remote Crews, Agent Harnesses, an…

05 **/config panel listing agents, MCP, Powers, Steering, Skills, Hooks**

NEW

70

Adds a `/config` command in V3 local or cloud sessions to view all configured agents, MCP servers, Powers, Steering, Skills, and Hooks in one terminal panel, with each item labelled as local, cloud, or both.

— lists every subsystem shown but not interaction detail

CLI 2.21.4: Session Dashboard, Configur…

06 **Terminal display settings in cli.json** NEW

70

Adds `chat.preserveScrollback` to control whether terminal history is preserved through redraws (defaults to `true`; set to `false` to restore clear-and-repaint behavior), and `chat.enableCustomSpinnerVerbs` with `chat.spinnerVerbs` in `cli.json` to customize busy indicators.

Disable the default history-preservation behaviour if your terminal emulator redraws incorrectly with scrollback retained.

```yaml
chat.preserveScrollback: false
```

Personalise the busy spinner verbs shown while Kiro processes a request.

```yaml
chat.enableCustomSpinnerVerbs: true
chat.spinnerVerbs:
  - Thinking
  - Reasoning
  - Planning
```

— exact keys, defaults, and runnable config examples given

CLI 2.21.4: Session Dashboard, Configur…

07 **Session browsing, resume, and search scope in CLI V3** NEW

65

V3 adds `kiro-cli chat --sessions` and a `/sessions` command to browse, search, resume, and clean up local and cloud past sessions. A new Session search setting in `/settings` lets you choose the indexing scope between Prompts only and Prompts and agent responses.

Enable full session search — including agent responses — so you can find past sessions by what Kiro answered, not just what you asked.

> 📍 In the CLI, open `/settings`, navigate to Session search, press Right or Enter, and select Prompts and agent responses.

— names commands and setting but not underlying search mechanics

CLI 2.21.4: Session Dashboard, Configur… CLI: Session Search Scope, V2 Harness F…

08 **--v2 flag for single-session harness override** NEW

65

Adds `--v2` flag to `kiro-cli chat` to run a single session on the V2 agent harness without overwriting the saved default.

Run a single chat session on the V2 agent harness without changing your saved default engine setting.

```
$ kiro-cli chat --v2
```

— exact flag with runnable example command CLI: Session Search Scope, V2 Harness F…

09 **Remote crew chat from sidebar** NEW 65

Adds `instances.enabled` config key and Settings > Developer > Feature Previews > Chat on a crew to start a chat on a connected remote crew from the sidebar without leaving the local UI.

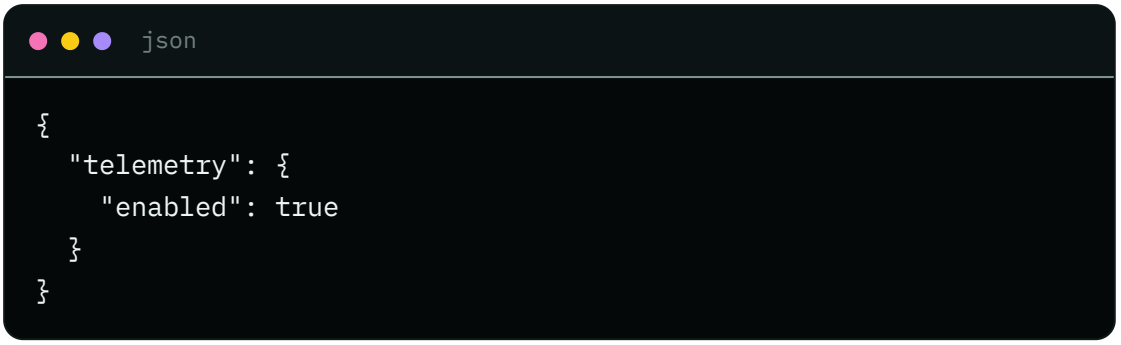
– config key and UI path given, no further mechanism

Crew: Remote Crews, Agent Harnesses, an…

10 **Developer Telemetry reporting** NEW 65

Adds `telemetry.enabled` config key to enable Developer Telemetry reporting of token use, spend, and latency across chats, jobs, subagents, workflows, and messaging channels.

Enable Developer Telemetry to track token spend and latency across all chats, subagents, and messaging channels.



– config key, scope of metrics, and example given

Crew: Remote Crews, Agent Harnesses, an…

11 **Two-hour default ceiling for unattended auto-run plans** 65

BREAKING

Adds `orchestrator.max_plan_duration_seconds` config key (default two hours) to control the ceiling for unattended auto-run plans; unattended auto-run plans now stop after two hours by default — raise the key or set it to `0` to restore unlimited run time.

– default change explained with exact mitigation key

Crew: Remote Crews, Agent Harnesses, an…

12 **Cloud Configuration Sync for CLI sessions** NEW 60

Enables Cloud Configuration Sync so that uploading a personal `.kiro` configuration in Kiro Web automatically applies cloud Steering, custom agents, Skills, Powers, and Hooks to new local CLI sessions without writing those files to the local directory.

– mechanism explained but no exact setup steps CLI 2.21.4: Session Dashboard, Configur…

13 **App manifest fields for jobs, commands, and session controls**

NEW 60

Adds `permissions.jobs`, `contributes.commands`, and `contributes.sessionControls` fields to app manifests, letting installed apps declare background jobs, Command Bar entries, and up to two live composer controls.

— three fields named, no example manifest shown Crew: Remote Crews, Agent Harnesses, an…

14 **AWS Control Overview and Files view** NEW 60

Adds AWS Control Overview showing account health, drive use, month-to-date spend, share links, and backup status, plus a Files view with preview, rename, search, upload, move, and delete.

— many concrete elements named but no navigation path given

Crew: Remote Crews, Agent Harnesses, an…

15 **Crew Members moved behind preview opt-in** BREAKING 60

Crew Members is now a preview opt-in; enable Settings > Developer > Feature Previews > Crew Members and Crew Mode to restore the Crew Members entry and new-crew-chat action.

— exact settings path given to restore prior behavior

Crew: Remote Crews, Agent Harnesses, an…

THINNER COVERAGE BELOW

16 **Per-session agent harness selection (Claude Code, Codex, KAS)** NEW 55

Adds agent harness selection — Claude Code, Codex, or KAS — per session via Settings > Developer > Agent Backend after enabling Developer Mode (preview).

— names three harnesses and UI path, no deeper mechanism

Crew: Remote Crews, Agent Harnesses, an…

17 **Agent avatar customization and Expressions pane** NEW 55

Adds agent avatar customization and an Expressions pane (working, done, failed states with face and sound) via Agent Capabilities > Agents > Avatar > Customize (preview).

— states and UI path named, preview-only feature Crew: Remote Crews, Agent Harnesses, an…

18 **Collapsed diff view with Plain diffs option** IMPROVED 55

Diffs now start folded into a file chip; Settings > Chat > Plain diffs enables raw patch text display.

— exact settings path given, minimal mechanism Crew: Remote Crews, Agent Harnesses, an…

19 **Session search by PR/issue badge and folder auto-tagging** IMPROVED 55

Adds session search by pull request, review, or issue badge, and automatic tagging of chats created in a folder; dormant sessions fold after 7 days by default (previously 2).

– concrete before/after numbers, limited mechanism detail

Crew: Remote Crews, Agent Harnesses, an…

20 **Python 3.12 minimum requirement** BREAKING 55

Python 3.12 is now the minimum required version; hosts on Python 3.10 or 3.11 must upgrade before installing or updating Kiro Crew.

– clear upgrade requirement, minimal further explanation

Crew: Remote Crews, Agent Harnesses, an…

21 **Credential-path protection change** BREAKING 55

Credential-path protection no longer relies on command-text inspection — verify that `agent.sandbox` in `config.json` is not set to `off` if you rely on the command gate.

– names config key to check, limited detail on new mechanism

Crew: Remote Crews, Agent Harnesses, an…

22 **Verbosity preview cycling in /verbosity settings** NEW 50

Adds a Verbosity preview inside `/verbosity` settings, where Ctrl+X cycles the preview through hidden, mini, and expanded states.

– brief description of preview cycle, no deeper mechanism

CLI: Session Search Scope, V2 Harness F…

23 **Session control enabled by default** IMPROVED 50

Adds `agent.session_control` config key to disable session control globally; session control is now on by default.

– states default change without explaining session control itself

Crew: Remote Crews, Agent Harnesses, an…

24 **Automatic knowledge folders removed** BREAKING 50

Automatic knowledge folders are removed; existing auto-registered folders remain pending until manually confirmed, and new folders must be added and confirmed before appearing in the Library.

– explains before/after but no exact remediation command

Crew: Remote Crews, Agent Harnesses, an…

25 **--system-python flag now explicit** BREAKING 50

The `--system-python` flag must now be passed explicitly when you want to use your system interpreter; the installer provisions its own pinned Python by default.

— exact flag named, minimal further explanation Crew: Remote Crews, Agent Harnesses, an…

26 **/usage command for plan and usage reporting** NEW 45

Adds `/usage` command that prints plan, limits, and usage percentages in non-interactive output.

Review your credit consumption and plan limits at any point during a non-interactive CLI session.

```
$ kiro-cli chat --sessions
```

— thin one-line description of command output CLI 2.21.4: Session Dashboard, Configur…

27 **kirocrew config defaults command** NEW 45

Adds `kirocrew config defaults` command to review and selectively adopt changed configuration defaults on upgrade.

Extend the hard ceiling for a long-running unattended plan, or remove it entirely for overnight automation runs.

```
$ kirocrew config defaults
```

— runnable command given but behavior thinly described

Crew: Remote Crews, Agent Harnesses, an…

28 **Crew members dispatch sub-agent workers** NEW 45

Crew members can now dispatch workers from their own threads on Claude and KAS backends, inheriting the creator's trust posture.

— capability stated without operational detail Crew: Remote Crews, Agent Harnesses, an…

29 **Apps Library Launchpad grid** NEW 45

Adds an Apps > Library Launchpad grid where installed apps can be opened, pinned, updated, disabled, or uninstalled.

— UI actions listed with navigation path, no further detail

Crew: Remote Crews, Agent Harnesses, an…

30 **Response indicator dot for agent messages** NEW 20

Adds a response indicator: a small dot in the agent's color marks each agent response.

— one-line cosmetic description only `CLI: Session Search Scope, V2 Harness F…`

└──▷ BREAKING ON UPGRADE

- ! Python 3.12 is now the minimum required version; hosts on Python 3.10 or 3.11 must upgrade before installing or updating Kiro Crew.
- ! Credential-path protection no longer relies on command-text inspection — verify that `agent.sandbox` in `config.json` is not set to `off` if you rely on the command gate.
- ! Unattended auto-run plans now stop after two hours by default; raise `orchestrator.max_plan_duration_seconds` or set it to `0` to restore unlimited run time.
- ! Crew Members is now a preview opt-in; enable Settings > Developer > Feature Previews > Crew Members and Crew Mode to restore the Crew Members entry and new-crew-chat action.
- ! Automatic knowledge folders are removed; existing auto-registered folders remain pending until manually confirmed, and new folders must be added and confirmed before appearing in the Library.
- ! Developing an app UI outside its install root from the dashboard or API is no longer supported; use `kirocrew app dev <name> --confirm-out-of-install-root` in the terminal instead.
- ! The `--system-python` flag must now be passed explicitly when you want to use your system interpreter; the installer provisions its own pinned Python by default.

WAS THIS USEFUL? 👍 👎

Replit Agent

Feedback ⓘ 1 RELEASE · 2026-09-13 CHANGELOG ↗

RANK

Replit Agent is an AI coding agent that builds, tests, and deploys applications in Replit.

> Replit Agent's single release this window added compliance-focused audit access, desktop MCP client connectivity, and managed PostgreSQL support for Databricks Apps.

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

01 **Compliance API prompt retrieval scope** NEW 75

Adds a `compliance:messages:read` API key scope to retrieve prompt text for `project.message_sent` audit events via the Compliance API, enabling security, DLP, SIEM, retention, and eDiscovery workflows.

— Names exact scope and event type but no endpoint path changelog-20260913-945ae0b2

02 **Lakebase managed PostgreSQL for Databricks Apps** NEW 62

Adds Lakebase as a managed production PostgreSQL database for Databricks Apps, with automatic schema migration on publish from Replit to a Databricks workspace.

— Describes mechanism (auto schema migration) but no config detail

changelog-20260913-945ae0b2

THINNER COVERAGE BELOW

03 **MCP desktop client support** NEW 50

Supports connecting Replit to any compatible desktop MCP client via the Replit MCP Server plugin, allowing users to create, find, inspect, update, and publish apps.

— Names the plugin and actions but no setup steps changelog-20260913-945ae0b2

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

Perplexity API

Feedback ⓘ 1 RELEASE · 2026-09-13 CHANGELOG ↗

RANK

Perplexity API Platform - build with the Router, Agent, Search, and Embeddings APIs. Real-time, web-wide research and Q&A capabilities for your products.

> Perplexity API added OAuth authentication for its remote MCP server and introduced a new model, google/gemini-3.8-flash, with promotional pricing through the end of 2026.

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

01 New model google/gemini-3.8-flash NEW

85

Adds `google/gemini-3.8-flash` as a callable model via the Perplexity API's `/chat/completions` endpoint, with promotional pricing through December 31, 2026: $0.75/M uncached-input tokens, $0.075/M cached-input tokens, and $3.75/M output and reasoning tokens.

Call google/gemini-3.8-flash via the Perplexity API to take advantage of promotional input pricing during the discount window.

```
$ curl https://api.perplexity.ai/chat/completions \
  -H 'Authorization: Bearer <your-api-key>' \
  -H 'Content-Type: application/json' \
  -d '{"model": "google/gemini-3.8-flash", "messages": [{"role":
"user", "content": "Summarize the latest threat intelligence
reports."}]}'
```

— Exact model name, pricing tiers, and a runnable curl example

changelog-20260913-0270d2d7

THINNER COVERAGE BELOW

02 OAuth support for remote Perplexity MCP server NEW

50

MCP clients (claude.ai, Claude Code, Cursor, VS Code) can now authenticate with a Perplexity account via OAuth instead of pasting an API key, with the billing organization chosen at sign-in.

— Names clients and mechanism but no config steps or endpoint changelog-20260913-0270d2d7

WAS THIS USEFUL?  

LMCache

Feedback ⓘ 1 RELEASE · 2026-09-12 NOTES ↗ RANK

LMCache reuses and shares language-model key-value caches across requests and serving instances to reduce repeated prefill work.

> LMCache v0.5.5 makes isolated CUDA IPC the default multiprocess memory-sharing path and promotes CacheBlend V3 to the sole blending engine, both breaking changes, alongside a durable pluggable coordinator, seven new accelerator/hardware backends, and richer multiprocess observability.

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

01 Coordinator gains durability, pluggability and inspection API

NEW 95

The coordinator gains a pluggable controller architecture (replacing static listing with discovery of views and controllers), durable state persisted to disk in named sections for the quota registry and LRU, a memory pressure API, per-key `ACCESS` count tracking in the key directory, a cache-event source abstraction, and namespace-scoped blend matching. It also exposes a `GET /cache/pins` endpoint for querying pinned cache entries and a `lmcache query coordinator` CLI subcommand for inspecting coordinator state.

Query the coordinator from the CLI to inspect its current state during a live MP deployment.

```
$ lmcache query coordinator
```

Retrieve all pinned cache entries from the coordinator to audit what is locked in memory.

```
$ curl -X GET http://<coordinator-host>:<port>/cache/pins
```

— Endpoint and CLI subcommand given with runnable examples. v0.5.5

02 Isolated CUDA IPC as default multiprocess backend

BREAKING 83

Adds `isolated_ipc` config key enabling raw CUDA IPC registration for KV caches, backed by a VMM CUDA IPC wrapper supporting `cuMemCreate` `-backed` allocations and a timeline-semaphore event IPC backend, removing the need for `hostIPC` or `/dev/shm` sharing in multiprocess deployments. `isolated_ipc` is now the default,

so existing configurations relying on `hostIPC` / `/dev/shm` sharing will need to be updated.

— Names config key, wrapper and removed dependencies but no example config. v0.5.5

---

03 **Seven new accelerator and hardware backends** NEW 75

Adds AWS Trainium (Neuron) backend support via the engine-driven path, Ascend NPU multiprocess format support with a generic engine-driven path and a pin-memory backend via AscendCL `aclrtHostRegister` for async MP transfers, native ATOM multiprocess integration with a configurable multiprocess key reader count, an Intel XPU wheel build and release workflow, a MooreThreads MUSA wheel workflow with a self-hosted smoke pipeline, a MetaX (MACA) self-compile build profile, CUDA 13.4.1 compatibility, and a rolling `nightly-rocm` wheel release including an ROCm 7.2 wheel for torch 2.10.

— Names every backend and wheel but no single usage command. v0.5.5

---

04 **No-warmup flag for benchmarking** NEW 70

Adds `--no-warmup` flag to `lmcache bench engine` to skip the warmup phase during benchmarking.

Run a CacheBlend engine benchmark without a warmup phase to get cold-start latency numbers.

```
$ lmcache bench engine --no-warmup
```

— Runnable benchmark command included as example. v0.5.5

---

05 **CacheBlend V3 becomes sole blend engine** BREAKING 67

`blend_v3` has been promoted to `blend` as the sole CacheBlend path, with the legacy blend implementation removed; configuration referencing the old `blend` path now uses the V3 engine. The engine also gains PD disaggregation support in `CacheBlendEngine` and linear hybrid model support.

— Names engine and new capabilities but no migration steps. v0.5.5

---

06 **New storage and P2P transfer backends** NEW 65

Adds a Phoenix backend for GDS L1 storage, uses the Mooncake Transfer Engine as the backend for P2P sharing, and batches `io_uring` `put_many` writes in bounded chunks for higher-throughput NVMe offload.

— Names backends and batching mechanism, no config example. v0.5.5

---

07 **Multiprocess transfer observability** IMPROVED 60

Adds L1/L2 hit attribution with early-exit reasons on lookup `END`, plus Prometheus counters for both; decomposes transfer timing into gather, DMA, and reserve phases; and adds inflight copy counters for multiprocess mode.

— Names metrics and phases but no query or dashboard path. v0.5.5

THINNER COVERAGE BELOW

08 **Dedup content flag for blending** NEW 55

Adds `--enable-dedup-content` flag to `lmcache` to skip duplicate fingerprint registration during blending.

— Flag named but no example invocation given. v0.5.5

09 **SGLang and MLA format support** NEW 50

Adds SGLang MLA format support in multiprocess mode, SGLang MUSA connectors with multiprocess MHA transfer, and MLA support with engine-driven transfer on CPU.

— Lists supported formats/connectors without usage detail. v0.5.5

10 **gRPC protobuf foundation for MP communication** NEW 25

Adds a gRPC protobuf foundation for multiprocess communication.

— Bare mention with no mechanism or usage detail. v0.5.5

└──▷ **BREAKING ON UPGRADE**

! Isolated IPC (`isolated_ipc`) is now the default for multiprocess deployments; `hostIPC` and `/dev/shm` sharing are no longer used and existing configurations that relied on them will need to be updated.

! `blend_v3` has been promoted to `blend` and the legacy blend implementation removed; any configuration or code referencing the old `blend` path will now use the V3 engine.

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 added a hosted shell server tool that lets any model execute isolated Linux commands and return their output, alongside a new Files API, both live in beta on the Responses and Messages APIs.

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

01 Shell server tool for sandboxed command execution NEW

85

A new hosted, isolated Linux container tool lets any model on OpenRouter run shell commands and get back `stdout`, `stderr`, and exit codes. It's available in beta on both the Responses and Messages APIs, invoked by passing `{"type": "shell"}` in the `tools` array.

Run an arbitrary shell command through a model on the Messages API and retrieve stdout — useful for automating code execution or environment probing via LLM.

```
$ curl https://openrouter.ai/api/v1/messages \
  -H 'Authorization: Bearer <your_api_key>' \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "<model_id>",
    "tools": [{"type": "shell"}],
    "messages": [{"role": "user", "content": "Run `uname -a` and
show me the output."}]
  }'
```

— Names mechanism, output fields, and includes a runnable curl example. snapshot-20260913

THINNER COVERAGE BELOW

02 Files API on Responses and Messages NEW

25

A new Files API is now available in beta on the Responses and Messages APIs, alongside the shell server tool.

— Named only, no endpoint, fields, or example given. snapshot-20260913

WAS THIS USEFUL? 👍 👎

Fireworks AI

Feedback ⓘ 1 RELEASE · 2026-09-13 CHANGELOG ↗

RANK

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

> Fireworks AI tightened deployment annotation permissions to a customer-managed namespace and shipped a training skill that lets coding agents plan and cost out fine-tuning jobs before spending.

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

01 Custom-prefixed deployment annotations enforced

BREAKING

90

`GetDeployment` and `ListDeployments` REST responses for regular account users now return only `custom/*` annotation entries, silently omitting any other keys. Writing a bare annotation key (e.g. `environment`) now returns `PERMISSION_DENIED`; keys must be prefixed as `custom/environment`. The new `firectl deployment tag set <deployment-id> environment=production` and `firectl deployment tag list <deployment-id>` subcommands let users set and verify these customer-managed tags, and the REST equivalent is a `PATCH https://api.fireworks.ai/v1/deployments/<deployment-id>` call with `{"annotations": {"custom/environment": "staging"}}`.

Tag a deployment with an environment label using the new firectl subcommand so you can filter deployments by lifecycle stage.

```
$ firectl deployment tag set <deployment-id>
environment=production
```

Write a customer-managed annotation via the REST API using the required `custom/` prefix to avoid a `PERMISSION_DENIED` error.

```
$ curl -X PATCH
https://api.fireworks.ai/v1/deployments/<deployment-id> \
  -H 'Authorization: Bearer <token>' \
  -H 'Content-Type: application/json' \
  -d '{"annotations": {"custom/environment": "staging"}}'
```

List all tags on a deployment to verify the `custom/` namespace migration is complete before the bare-key cutover.

```
$ firectl deployment tag list <deployment-id>
```

— Names exact endpoints, error code, prefix, and firectl commands with runnable examples.

changelog-20260913-3cefaf20

THINNER COVERAGE BELOW

02 **Fireworks training skill for coding agents** NEW 46

A new Fireworks training skill works with Claude Code, Cursor, Codex, and other compatible coding agents, enabling plain-language training job planning, cost estimation, and approval gating before spend is committed.

— Names supported agents and capabilities but gives no command or setup steps.

changelog-20260913-3cefaf20

└──▷ **BREAKING ON UPGRADE**

! Bare annotation keys (e.g. `environment`) now return HTTP `PERMISSION_DENIED` on REST writes; keys must be prefixed with `custom/` (e.g. `custom/environment`).

! `GetDeployment` and `ListDeployments` no longer return non-`custom/*` annotation keys for regular account users; clients reading bare keys will silently receive no data for those entries.

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

KTransformers

Feedback ⓘ 1 RELEASE · 2026-09-13 NOTES ↗ RANK

KTransformers provides kernels for language-model inference and fine-tuning with heterogeneous CPU and GPU execution.

> KTransformers v0.7.0.post4 adds a full Kimi K2.5 LoRA fine-tuning workflow with checkpoint resume and SGLang adapter loading, extends model coverage to Qwen3-30B-A3B and DeepSeek-V3.1, and publishes its five core wheels to public PyPI with verified checksums.

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

01 **Core wheels published to public PyPI** NEW 80

Publishes all five core wheels — `ktransformers`, `kt-kernel`, `sglang-kt` 0.7.0.post4, `transformers-kt` 5.6.0.post5, `accelerate-kt` 1.14.0.post3 — to public PyPI with verified SHA256 values, installable via the `kimi-k25-post4-user-kit.tar.gz` archive checked against `SHA256SUMS.txt`.

Install the full v0.7.0.post4 stack from PyPI after verifying the user kit archive.

```
$ wget https://github.com/kvcache-ai/ktransformers/releases/download/v0.7.0.post4/kimi-k25-post4-user-kit.tar.gz
wget https://github.com/kvcache-ai/ktransformers/releases/download/v0.7.0.post4/SHA256SUMS.txt
sha256sum -c SHA256SUMS.txt
tar -xzf kimi-k25-post4-user-kit.tar.gz
cd kimi-k25-post4 && cat README.md
```

— Exact package names, versions and a runnable install command v0.7.0.post4

THINNER COVERAGE BELOW

02 **Kimi K2.5 LoRA training and fine-tuning workflow** NEW 53

Adds Kimi K2.5 text LoRA training support, validated on 8 RTX 5090 GPUs, plus checkpoint resume for exact continuation of Kimi K2.5 training runs and independent SGLang adapter loading for Kimi K2.5 fine-tuned adapters.

— Names GPU count and mechanism but no flags or commands given v0.7.0.post4

03 **New model support: Qwen3-30B-A3B and DeepSeek-V3.1** NEW 40

Adds Qwen3-30B-A3B BF16 training and base inference support, and DeepSeek-V3.1 native FP8 training support.

— Names models and precision formats, no usage detail `v0.7.0.post4`

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

Pinecone

Feedback ⓘ 1 RELEASE · SEEN 2026-09-13 NOTES ↗ RANK

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

> Pinecone's biggest addition this window is a new Documents API built on schema-based indexes — a breaking change requiring migration — alongside a broad set of administrative endpoints for projects, organizations, API keys, invites, service accounts and role bindings, plus a metadata indexing control to cut query overhead.

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

01 **Metadata indexing configuration for filters** NEW 45

Adds metadata indexing configuration that lets users limit indexed fields to only those used in filters, reducing index build and query execution overhead.

— Describes benefit but no config key or endpoint named `snapshot-20260913`

02 **Role binding creation endpoint** NEW 42

Adds a Create role binding endpoint to grant a role to a principal at organization or project scope.

— Single endpoint named, no further mechanism or scope detail `snapshot-20260913`

└──▷ BREAKING ON UPGRADE

! API version 2026-07 introduces schema-based indexes via the Documents API; existing code must be evaluated for compatibility and migrated according to the adoption guide.

WAS THIS USEFUL? 👍 👎

timgordontg engrim

Feedback ⓘ 1 RELEASE · 2026-09-13

NOTES ↗ RANK

The Universal Cross-Model Episodic Memory Standard. Local-first, project-scoped SQLite memory engine for Google Antigravity, Claude Code, Cursor, Windsurf, and Codex.

> engrim v1.4.3 makes its core MCP tools introspectable by declaring explicit output schemas for structured clients.

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

01 **Output schemas for core MCP tools** IMPROVED 66

Adds explicit `outputSchema` declarations to `engrim_recall`, `engrim_context`, `engrim_add`, and `engrim_review` in `src/engrim/mcp_server.py`, enabling MCP clients to introspect structured handler payloads.

— Names exact tools and file but no behavior change detail v1.4.3

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ

2 RELEASES · 2026-08-20 → 2026-09-13 | CHANGELOG ↗ | RANK

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

> LangSmith's biggest window items are a new `/v2/threads/stats` analytics endpoint, a public `POST /v2/datasets/<id>/experiment-runs` experiment-comparison API, and expanded Fleet agent sandbox/OAuth/Databricks support, alongside LLM Gateway fallback and pricing metadata improvements, dataset-splits and evaluator UX work, and several deprecations of legacy feedback and dataset-comparison endpoints.

└──▷ WHAT SHIPPED · 29 FEATURES most completely described first ? what's the number?

01 **Fleet MCP caller header forwarding** NEW — 85

Adds `FLEET_MCP_FORWARD_CALLER_HEADERS=true` environment variable to forward a REST caller's `X-Fleet-Forward-*` headers to custom MCP servers an agent calls, enabling per-invocation context such as end-user identity.

Enable forwarding of caller-supplied headers to custom MCP servers so a policy gateway can enforce per-invocation identity checks.

```env
FLEET_MCP_FORWARD_CALLER_HEADERS=true
```

— Exact env var, header pattern, and example config changelog-20260913-009c9b77

02 **Thread and trace statistics API** NEW — 80

Adds `GET /v2/threads/stats` endpoint to retrieve thread and trace counts, latency, token, cost, and feedback statistics for a tracing project; available on self-hosted deployments backed by ClickHouse, including when SmithDB queries are disabled.

— Named endpoint and stats scope, no example call given changelog-20260913-009c9b77

03 **License expiration warning window** NEW — 75

Adds `LICENSE_EXPIRATION_WARNING_DAYS` environment variable for self-hosted operators to configure how many days before license expiration the warning banner appears, defaulting to 7 days.

Set a warning banner lead time of 30 days before license expiration on a self-hosted LangSmith deployment.

```env
LICENSE_EXPIRATION_WARNING_DAYS=30
```

— Exact env var, default, and runnable example changelog-20260913-009c9b77

04 **Fleet agent-scoped sandbox provisioning** NEW 75

Fleet now provisions agent-scoped sandboxes during agent creation and returns the sandbox ID and status in the response; agents include a server-generated slug for addressing agent-scoped sandboxes through the sandbox API; and clients can activate the sandbox referenced by a thread and receive its ready status.

— Describes provisioning, addressing, and activation mechanics

changelog-20260913-009c9b77

05 **Gateway-backed prompt manifests** IMPROVED 75

Prompts saved with a LangSmith Gateway model now use a standard `ChatOpenAI` manifest with the unified Gateway URL and a `LANGSMITH_GATEWAY_API_KEY` secret reference, so they can be pulled and run with standard LangChain SDKs.

— Names manifest type, secret key, and resulting SDK compatibility

changelog-20260913-009c9b77

06 **Public experiment comparison API** NEW 70

Introduces `POST /v2/datasets/<id>/experiment-runs` as the supported public API for paginated experiment comparison.

— Named endpoint with method and path, no example call snapshot-20260913

07 **Thread evaluator testing via validate endpoint** IMPROVED 70

Extends `POST /runs/rules/validate` endpoint to support thread evaluators via `test_thread_id` and `session_id` parameters, enabling testing of a multi-turn evaluator against a real conversation before saving, avoiding a gap where an old evaluator rule is deleted before the new one is confirmed.

Update an existing code evaluator rule in place after editing the implementation, avoiding a gap where the old rule is deleted before the new one is confirmed.

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

— Named endpoint and parameters with related workflow example snapshot-20260913

08 **Fleet Databricks model configurations** NEW 65

Fleet agents now support saved Databricks model configurations, using configured workspace URL, serving endpoint, and workspace credentials via Databricks Model Serving or AI Gateway routes.

— Names concrete config surface, no example given `changelog-20260913-009c9b77`

09 **LLM Gateway fallback timeout and context retention** IMPROVED

65

LLM Gateway fallback chains can now advance to the next model when a candidate does not return response headers within a configured timeout, while active response streams continue after headers arrive; gateway requests authenticated with a session or OAuth bearer token retain authentication and workspace context across internal routing, including model fallbacks.

— Explains fallback mechanism and auth propagation, no config key named `changelog-20260913-009c9b77`

10 **Thread-to-dataset and annotation queue actions** NEW

65

The thread header now has Add to dataset and Add to annotation queue buttons in every thread view (Messages, Turns, and Details), so a whole thread can be added without selecting a run; the add-to-dataset popup for threads also offers a New Dataset option that creates the dataset empty and adds the selected threads to it.

— Exact UI paths and views named `changelog-20260913-009c9b77`

11 **Annotation queue item batch limit** BREAKING

65

`POST /annotation-queues/<id>/items` now accepts at most 200 items per request and returns a clear validation error when the limit is exceeded.

Add a batch of conversation threads to an annotation queue for human review, mixing thread and run items in a single request.

```
$ curl -X POST 'https://api.smith.langchain.com/annotation-
queues/<queue_id>/items' \
  -H 'Content-Type: application/json' \
  -H 'x-api-key: <LANGSMITH_API_KEY>' \
  -d '{"items": [{"item_type": "THREAD", "thread_id": "
<thread_id>", "session_id": "<session_id>"}, {"item_type": "RUN",
"run_id": "<run_id>"}]}'
```

— Named endpoint, exact limit, and runnable example request `snapshot-20260913`

12 **Bedrock Playground timeout and retry controls** IMPROVED

60

Amazon Bedrock models in the Playground now expose request timeout and max retries options, removing the previous 60-second AWS client default cutoff for long-running agent calls.

— Names before/after limit but no UI path or values changelog-20260913-009c9b77

13 **Managed Deep Agent channel failure visibility** IMPROVED

60

Managed Deep Agent channel reply failures (dropped delivery, missing bot token, ambiguous route) now surface in `mda logs` and the deployment logs UI instead of failing silently.

— Names failure modes and log surfaces, no example log output changelog-20260913-009c9b77

14 **Dataset splits bulk management and display** NEW 60

Enables bulk selection of experiment rows to add, replace, or remove dataset splits in one action, or copy selected examples to another dataset; each example's dataset splits are now surfaced as chips in the dataset Examples table, with a clickable +N overflow menu.

— Describes UI actions across two related additions snapshot-20260913

THINNER COVERAGE BELOW

15 **Provider metadata field on model configs** NEW 55

Adds `ls_provider` field on OpenAI-compatible model configurations to populate trace metadata and match custom models to provider-specific pricing through the LLM Gateway.

— Named field but no usage walkthrough changelog-20260913-009c9b77

16 **OAuth authorization session owners and branding** NEW

55

OAuth authorization sessions now accept `owner_type` and `owner_id` fields, allowing a human to authorize a managed credential for either a user or an agent, and now also accept workspace owners so one managed OAuth credential can be shared across a workspace; agent OAuth connections now show a provider-branded result page after authorization.

— Named fields, mechanism thin on how sharing is enforced changelog-20260913-009c9b77

17 **Download permission required for dataset exports** BREAKING

55

Dataset CSV and JSONL exports now require both dataset read and download permissions; direct API requests without download permission return a forbidden

response.

— Clear before/after but no migration example changelog-20260913-009c9b77

18 **On-demand IAM role assumption for private ECR** IMPROVED

50

Commercial private ECR registries can now assume an AWS IAM role on demand, enabling repository and tag search while avoiding expiring saved authorization tokens.

— Explains mechanism, lacks setup steps changelog-20260913-009c9b77

19 **Experiment and dataset table rendering improvements** IMPROVED

50

Experiment and dataset tables now truncate long input and reference-output text and show detected base64 images as small thumbnails with a delayed larger preview; tables also defer full payload rendering and output diff preparation until those views are requested, improving responsiveness for large agent trajectory runs.

— Describes UI/performance mechanism, no config surface snapshot-20260913

20 **Server-sent events for thread queries** NEW

45

Thread query clients can request server-sent events to receive results progressively while preserving cursor pagination.

— Behavior named but no request example changelog-20260913-009c9b77

21 **Claude Fable 5.1 model availability** NEW

45

Claude Fable 5.1 is now available in the Anthropic, Bedrock, and Vertex AI model selectors in the Playground.

— Names model and selectors, no further detail changelog-20260913-009c9b77

22 **MCP connector Client ID Metadata Documents** IMPROVED

45

LangSmith MCP connectors now use Client ID Metadata Documents when supported, with dynamic client registration retained as a compatibility fallback.

— Names mechanism but no configuration steps changelog-20260913-009c9b77

23 **Dedicated gRPC service for trajectory batch reads** IMPROVED

45

Trajectory batch reads now run on a dedicated in-cluster gRPC service instead of the `smith-go` monolith, so load from trajectory evaluations no longer competes with UI and API traffic.

— Explains infra change, nothing for reader to act on changelog-20260913-009c9b77

24 **Application resource tagging** NEW 45

Workspace editors can now assign resources to an Application tag value from Settings > Resource Tags.

— Names UI path, thin on scope `changelog-20260913-009c9b77`

25 **IssueBench Opus 5 regression task** NEW 40

IssueBench now includes a benchmark task that compares pre- and post-Opus 5 agent traces for cost and trajectory regressions.

— Names the task, no usage instructions `changelog-20260913-009c9b77`

26 **Context Hub markdown Source mode fidelity** IMPROVED 40

Context Hub now opens markdown files in Source mode on the exact bytes stored in the repo, and saving a change no longer rewrites files you did not edit.

— Clear before/after, no named config surface `changelog-20260913-009c9b77`

27 **Evaluator error recording on timeout or failure** IMPROVED 40

Custom code evaluators that time out or fail on a run now record an error on that run instead of silently leaving it without feedback.

— Behavior described, no surface or example named `snapshot-20260913`

28 **Workspace switcher by ID** IMPROVED 30

The workspace switcher now accepts workspace IDs in addition to workspace names.

— Single-line description with no further mechanism `changelog-20260913-009c9b77`

29 **Evaluator spend chart axis formatting** IMPROVED 25

Evaluator spend charts now abbreviate y-axis amounts of $1,000 or more.

— Minor cosmetic change, thin description `snapshot-20260913`

└──▷ BREAKING ON UPGRADE

! Dataset CSV and JSONL exports now require both dataset read AND download permissions; API requests that previously succeeded with only read permission now return a forbidden response.

! The legacy feedback formula endpoints (`POST/GET /feedback/formulas` and `GET/PUT/DELETE /feedback/formulas/{feedback_formula_id}`) are deprecated and scheduled for removal on 2026-08-20; migrate existing feedback formulas to composite evaluators.

! Legacy dataset comparison helpers are removed from the public OpenAPI spec and generated SDKs (existing HTTP routes continue to work for LangSmith UI clients).

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ 2 RELEASES · 2026-09-01 → 2026-09-13

CHANGELOG ↗ RANK

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

> Braintrust's headline addition is Patterns, a Loop-powered feature that automatically surfaces recurring trace problems and lets teams turn findings into scorers or prompts, alongside a move of Loop itself onto a Braintrust-managed runtime with built-in models. The window also expands CLI tracing for coding agents (Grok, Antigravity, opencode, pi, Claude Code, Codex) with several breaking changes, adds Metric columnstore, Dashboard sections, GLM-5.3 Flash, observability templates, a Rust SDK CA-bundle option, and Go/Java SDK and CLI subcommand additions.

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

01 CLI tracing integrations for coding agents BREAKING

90

The `bt` CLI (v0.19.3+) can trace Grok and Google Antigravity coding sessions as Braintrust traces, capturing prompts, model calls, and tool activity without interrupting the agent if tracing fails. The `bt trace enable opencode` and `bt trace enable pi` plugins now install v2 of their tracing integrations (see the v0.19.3 migration instructions) and no longer read environment-variable overrides for tracing enablement, profile, organization, project, or metadata — settings must be saved via those commands instead. Normal Claude Code and Codex sessions no longer read `TRACE_TO_BRAINTRUST` or `BRAINTRUST_ADDITIONAL_METADATA` from the environment, `bt trace run codex` now rejects `--dangerously-bypass-hook-trust` to prevent duplicate Braintrust hooks, and the Braintrust Claude and Codex marketplaces no longer distribute the MCP-specific plugin (or, for Codex, its bundled routing skill) — configure Braintrust MCP directly in Claude/Codex or use a remote connector in Cowork.

— Names every affected command, flag, and env var with migration pointer

snapshot-20260913 changelog-20260913-72c92428

02 Observability templates via CLI NEW

85

Adds `bt observability template export` and `bt observability template apply` commands to export one project's observability setup as a reusable JSON template and apply it to another project, avoiding manual facet and automation rebuilding.

Export one project's observability setup as a reusable template and apply it to a new project, avoiding manual facet and automation rebuilding.

```
$ bt observability template export --project my-agent-prod >
template.json
bt observability template apply --project my-agent-staging
template.json
```

— Runnable export/apply commands given directly `snapshot-20260913` `changelog-20260913-72c92428`

03 **Java SDK concurrency and OTel tuning controls** BREAKING

85

Adds .maxConcurrency(int) on `Devserver.Builder` and .executor(Executor) for supplying a custom `Executor` in the Java SDK, plus `BRAINTRUST_DEFAULT_MAX_CONCURRENCY`, `BRAINTRUST_OTEL_MAX_QUEUE_SIZE`, `BRAINTRUST_OTEL_MAX_EXPORT_BATCH_SIZE`, and `BRAINTRUST_OTEL_EXPORT_INTERVAL_MILLIS` environment variables for global concurrency and OTel batch tuning defaults. As a result, `Task` and `Classifier` implementations in the Java SDK must now be thread-safe due to concurrent eval case execution.

— Names every method and env var plus the resulting requirement `changelog-20260913-72c92428`

04 **bt sync pull default history window change** BREAKING

80

`bt sync pull` now includes all matching experiment and dataset history by default; pass `--window` or set `BT_SYNC_WINDOW` to restore a bounded window. Project logs retain the default three-day window — see the v0.19.2 migration instructions.

— Names the flag and env var needed to restore old behavior `snapshot-20260913`

05 **Loop moves to managed runtime** IMPROVED

75

Loop now runs in a Braintrust-managed runtime instead of the browser, enabling longer investigations and persistent threads, and adding the ability to create and edit prompts, scorers, datasets, facets, custom views, dashboards, and automations directly from Loop; it now includes built-in GPT-5.6 Sol, GPT-5.6 Terra, and GPT-5.6 Luna models so no AI provider configuration is required.

— Names models and new capabilities but no runnable entry point `snapshot-20260913`

06 **Custom CA bundle for Rust SDK uploads** NEW

75

Adds support for stamping traces with a private CA bundle via `BRAINTRUST_CUSTOM_CA_BUNDLE`, so trace uploads from the Rust SDK are not

rejected when running behind a corporate TLS proxy.

Stamp traces with your private CA bundle when the Rust SDK is behind a corporate TLS proxy, so uploads are not rejected.

```
$ BRAINTRUST_CUSTOM_CA_BUNDLE="$(cat /etc/ssl/certs/corp-ca.pem)"
bt trace run pi
```

— Comes with a runnable example command using the env var `snapshot-20260913`

07 **Patterns automated trace pattern detection** NEW 65

Introduces Patterns, a scheduled Loop-powered feature that searches traces for recurring problems and trends, saves each finding with supporting traces and suggested fixes, and lets users create a scorer, continue the investigation in Loop, or copy the finding as a prompt.

— Clear mechanism and workflow but no command or exact UI path given `snapshot-20260913`

08 **New bt subcommands for environments, prompts, functions** NEW 65

Adds `bt environments` commands to manage deployment environments, `bt prompts` commands for prompt version listing and environment assignment, and `bt functions view` for environment-based version lookup.

— Names three subcommands but no example invocation `changelog-20260913-72c92428`

09 **Dashboard sections for chart grouping** NEW 60

Adds Dashboard sections, allowing charts to be grouped into named sections that can be reordered, duplicated, and collapsed, with collapsed state persisted per browser.

— Describes behavior but no navigation or config detail `snapshot-20260913`

10 **MCP org_name parameter for name lookups** IMPROVED 60

Adds `org_name` parameter to MCP tools for disambiguating name lookups across organizations, with pattern, prompt, and scorer links now resolving within their containing projects.

— Names the parameter but no example call shown `changelog-20260913-72c92428`

11 **Prompt support in Go SDK** NEW 60

Adds prompt support to the Go SDK: load a Braintrust prompt, render variables locally, call any LLM client, and link rendered calls back to their prompt in Braintrust traces.

— Describes workflow but no code sample or function name `changelog-20260913-72c92428`

THINNER COVERAGE BELOW

12 **Vercel AI SDK v6 tool call metadata** IMPROVED 30

Adds Vercel AI SDK v6 tool call `toolCallId` field in span metadata.

— Bare field name with no usage context changelog-20260913-72c92428

└──▷ ALSO FROM THESE RELEASES

List all project details as JSON for use in a CI automation script, without opening a browser.

```
$ bt projects view --json
```

└──▷ BREAKING ON UPGRADE

! The `bt trace enable opencode` and `bt trace enable pi` plugins no longer read environment-variable overrides for tracing enablement, profile, organization, project, or metadata — settings must be saved with `bt trace enable opencode` or `bt trace enable pi`.

! Normal Claude Code and Codex sessions no longer read `TRACE_TO_BRAINTRUST` or `BRAINTRUST_ADDITIONAL_METADATA` from the environment.

! `bt trace run codex` rejects `--dangerously-bypass-hook-trust` to prevent duplicate Braintrust hooks.

! The Braintrust Claude marketplace no longer distributes the MCP-specific plugin; configure Braintrust MCP directly in Claude or use a remote connector in Cowork.

! The Braintrust Codex marketplace no longer distributes the MCP-specific plugin and its bundled routing skill; configure Braintrust MCP directly in Codex.

! `Task` and `Classifier` implementations in the Java SDK must now be thread-safe due to concurrent eval case execution.

! `bt trace enable opencode` and `bt trace enable pi` now install v2 of their respective tracing integrations — see the v0.19.3 migration instructions before upgrading.

! `bt sync pull` now includes all matching experiment and dataset history by default; pass `--window` or set `BT_SYNC_WINDOW` to restore a bounded window. Project logs retain the default three-day window — see the v0.19.2 migration instructions.

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Anthropic

Feedback ⓘ 1 RELEASE · 2026-09-13 CHANGELOG ↗ RANK

Anthropic provides Claude AI models and APIs for building applications that generate, analyze, and automate work.

> Anthropic added human-in-the-loop permission policies to Claude Managed Agents, plus two new CLI workflows for managing and monitoring those agents: declarative infra-as-code apply and live session control.

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

01 `ant apply` for infra-as-code agent management NEW

75

Declare all your agents and supporting resources as files and apply them in one command (`ant apply`) so CI always updates the same resources instead of creating duplicates; changes are tracked by committing an updated `claude-lock.json` .

Declare all your agents and supporting resources as files and apply them in one command so CI always updates the same resources instead of creating duplicates.

```
$ ant apply && git add claude-lock.json && git commit -m 'chore: update claude-lock.json'
```

— Runnable command and named lock file shown end to end. changelog-20260913-a850a3bd

02 Permission policies for Claude Managed Agents NEW

70

Adds permission policies to Claude Managed Agents that evaluate each `agent.tool_use` and `agent.mcp_tool_use` call, letting it run, be denied, or pause for human approval; the outcome is reported via an `evaluation` field and an `evaluated_permission` field on each event.

— Names event types and result fields but no config command shown.

changelog-20260913-a850a3bd

WAS THIS USEFUL? 👍 👎

OpenAI

Feedback ⓘ 1 RELEASE · 2026-09-13 CHANGELOG ↗ RANK

OpenAI provides AI models and APIs for building applications that generate and reason with text, images, and audio.

> OpenAI shipped a real-time voice API, two new high-quality image models, a specialized research model, generally available prompt cache diagnostics, and API key expiration controls in this release wave.

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

01 **New xhigh-quality image models** NEW 85

Adds `gpt-image-2.5-sunburst` and `gpt-image-2.5-flare` models to `v1/images` and the Responses API image generation tool, both supporting a new `xhigh` quality setting priced at GPT Image 2 token rates.

Generate a high-fidelity image at `xhigh` quality using the precision-editing model `gpt-image-2.5-sunburst` via the Image API.

```
$ curl https://api.openai.com/v1/images/generations \
  -H 'Authorization: Bearer $OPENAI_API_KEY' \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "gpt-image-2.5-sunburst",
    "prompt": "A photorealistic render of a server rack in a dark data center",
    "quality": "xhigh",
    "n": 1
  }'
```

— Runnable curl example included with named endpoint and models

changelog-20260913-145f0fd5

02 **GPT-Live 1 real-time voice sessions** NEW 80

Adds the `gpt-live-1` model via `v1/live/sessions` for full-duplex voice conversations, supporting Responses delegation (OpenAI model) or client delegation (your own backend), billed at $0.05/min per second.

— Names endpoint and pricing but no runnable example given changelog-20260913-145f0fd5

03 **Prompt Cache Diagnostics generally available** IMPROVED 60

Makes Prompt Cache Diagnostics generally available in the Responses API for `gpt-5.6` and later, enabling cache-miss identification and troubleshooting guidance against a

previous response.

— Explains mechanism and model scope, no concrete command shown

changelog-20260913-145f0fd5

THINNER COVERAGE BELOW

04 **GPT-Rosalind research model** NEW 55

Adds the `gpt-rosalind-research` model via the trusted-access program for approved life sciences research, priced at $5/1M input tokens, $0.50/1M cached input tokens, and $25/1M output tokens.

— Pricing and access program named but no usage path changelog-20260913-145f0fd5

05 **API key expiration controls** NEW 50

Enables expiration dates on project API keys, with organization- and project-level enforcement of maximum key lifetime configurable in Platform settings.

— Clear UI starting point but no exact settings path given changelog-20260913-145f0fd5

06 **Agents API** NEW 10

OpenAI's changelog summary mentions an Agents API shipping in this release wave, but no further detail on its endpoints, capabilities, or behavior is provided.

— Mentioned only by name with no supporting detail changelog-20260913-145f0fd5

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · September 13, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue