NOW AVAILABLE The 512GB Mac Studio is on sale now, shipping February to March. See every configuration.

THE AI TOOLCHAIN

NO. 086

# Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED NOVEMBER 7, 2026 · EVERY DAY

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

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

▸ // HOW THIS ISSUE IS MADE SHOW

VIEW

ISSUE VIEW

**scroll**



SCROLL | EXPAND IN PLACE | PREVIEW PANE

INFORMATION DENSITY

**everything**



BRIEFED | EVERYTHING

Every tool in the issue, in full.

all 334 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 84 of 88 tools

▾ // THE BRIEF

HIDE SAME ISSUE, SAME PROMPT, TWO WRITERS: Opus 5 | GPT-5.5

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

vim ? help

Three unrelated things stand out. NeMo RL's v0.8.0 decouples generation from training so rollouts stream into the learner instead of blocking it, and adds TensorRT-LLM and Dynamo rollout engines. Google ADK adds multi-run evals with configurable parallelism and semantic matching on tool arguments. Kiro's Compaction Hooks let you intercept what gets thrown away when a conversation is summarized.

EVALUATE

**Run an agent eval several times over and stop failing it on argument phrasing**

ADK v2.14.0 runs each eval case multiple times with configurable parallelism, which is the only honest way to score a stochastic agent, and matches tool arguments semantically so a correct call phrased differently stops counting as a miss. Inspect AI 0.3.288 adds transcript events for agent swarms and per-agent sandbox user isolation, so a multi-agent failure can be attributed to one agent rather than read as one undifferentiated log.

Google Agent Development Kit · Inspect AI

BUILD

**Keep the RL learner busy while rollouts are still generating**

The v0.8.0 Single Controller async runtime streams training alongside generation with pluggable samplers and fault tolerance, so the GPUs doing updates no longer idle waiting for a batch of completions to finish. Faster NCCL/NIXL weight-refit transports cut the other half of that stall — pushing updated policy weights back to the rollout workers — and LoRA plus NVFP4/W4A4 training on Blackwell lower what a run costs to stand up.

NeMo RL

BUILD

**Decide what survives a context compaction instead of letting the agent guess**

Kiro's Compaction Hooks fire across CLI and IDE when a conversation is summarized, so a team can pin the facts an agent keeps losing on long tasks rather than rediscovering them; persistent Memory now readable by sub-agents closes the same gap across sessions. OpenCode's percentage-based compaction buffer means one threshold works whether

> the model has a small or large context window, and git worktree isolation keeps scripted runs from stepping on each other's working tree.
>
> Kiro · OpenCode

> OPERATE
>
> **See which provider and route a traced call actually took**
>
> Gateway metadata now lands in traces, so when a call is slow or wrong you can tell whether routing, the provider, or the prompt is at fault instead of inferring it. IAM role authentication for Amazon Bedrock removes long-lived keys from gateway config.
>
> LangSmith

Does **Opus 5** read better? 👍 👎

## OFFENSIVE SECURITY

◆ Exploitation & C2

# xalgorix

Feedback ⓘ v4.6.182 NOTES ↗ RANK

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

> xalgorix shipped three breaking-by-default scanning changes this window — bounded working context that archives old tool output behind `read_tool_output`, port-pinned scan boundaries, epoch-based coverage convergence, and narrowed single-target scan scope — alongside improved LLM gateway provider routing and diagnostics.

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

## 01 Bounded working context by default

BREAKING 95

xalgorix now enables bounded working context by default, archiving old tool output and serving it on demand via `read_tool_output`. Set `XALGORIX_BOUNDED_CONTEXT=false` to restore the previous behaviour of resending full tool output on every iteration, or tune `XALGORIX_TOOL_ARCHIVE_ACTIVE_WINDOW` and `XALGORIX_TOOL_ARCHIVE_MIN_BYTES` to control the active window size and the minimum byte threshold before old results are compacted.

Disable bounded context to restore the previous behaviour of resending full tool output on every iteration.

```
$ XALGORIX_BOUNDED_CONTEXT=false xalgorix scan <target>
```

Tune the active window and minimum byte threshold so only large, old results are compacted — useful when scans produce many small outputs you want to keep verbatim.

```shell
XALGORIX_BOUNDED_CONTEXT=true
XALGORIX_TOOL_ARCHIVE_ACTIVE_WINDOW=10
XALGORIX_TOOL_ARCHIVE_MIN_BYTES=4096
```

— Full mechanism, env vars and tuning example given. v4.6.182

## 02 Port-pinned scan boundaries

BREAKING 90

Including a port in the target URL now restricts the entire scan to that port: requests, port scans, and findings on any other port of the same host are refused, and service

enumeration must name only the pinned port with `-p <port>`, e.g. `xalgorix scan https://example.com:9000 -p 9000`. Previously a scan of a target like `https://example.com:9000` could still discover or report findings on other ports such as `:81` or `:5013`.

Scope a scan strictly to a non-standard port so findings on other services on the same host are never reported.

```
$ xalgorix scan https://example.com:9000 -p 9000
```

— Exact command, flag and before/after given. v4.6.181

---

03 **LLM gateway provider routing and diagnostics** IMPROVED

90

Gateway and inference host providers such as OpenRouter, Groq, Together, vLLM, and LM Studio now receive the full model ID unchanged (e.g. `google/gemini-3-pro`, `qwen/qwen3-32b`) when `XALGORIX_LLM_PROVIDER` is set, instead of having the prefix stripped before the request is forwarded. Built-in providers with no legacy default, including OpenRouter and Together, now fall back to their catalog base URL when `XALGORIX_API_BASE` is unset instead of a generic default host, and xalgorix logs a one-time message when a provider setting overrides a model ID prefix so stale `XALGORIX_LLM_PROVIDER` or `XALGORIX_API_BASE` settings are easy to spot.

Pin the base URL explicitly in a Docker or CI environment to override the catalog default for a self-hosted gateway like vLLM.

```ini
XALGORIX_LLM_PROVIDER=vllm
XALGORIX_API_BASE=http://vllm-host:8000/v1
XALGORIX_API_KEY=<your-key>
```

— Names providers, env vars, model IDs with config sample. v4.6.180

---

04 **Epoch-based coverage convergence** BREAKING

80

`XALGORIX_EPOCH_COVERAGE` is now enabled by default so scans close on large or growing attack surfaces by measuring completion against a frozen surface snapshot that advances in bounded epochs. Set `XALGORIX_EPOCH_COVERAGE=false` to turn it off and restore the previous unbounded completion behaviour.

Opt out of epoch-based coverage convergence when you need the previous unbounded completion behaviour.

```
$ XALGORIX_EPOCH_COVERAGE=false xalgorix scan https://example.com
```

— Mechanism and env var given, no epoch size numbers. v4.6.181

THINNER COVERAGE BELOW

05 **Narrowed single-target scan scope** BREAKING 55

Single-target scans now cover only the configured host — subdomain discovery, takeover checks, and sibling subdomain content discovery no longer apply and are silently skipped in single mode, whereas previously these ran against a registrable domain target.

— Behaviour change described, no opt-out flag given. v4.6.181

└──▷ **BREAKING ON UPGRADE**

! `XALGORIX_BOUNDED_CONTEXT` now defaults to `true`; agents that relied on full tool output being resent verbatim on every iteration will instead receive compact retrieval stubs for older results.

! `XALGORIX_EPOCH_COVERAGE` now defaults to `true`; deployments that relied on the previous off-by-default behaviour must explicitly set `XALGORIX_EPOCH_COVERAGE=false` to preserve it.

! A port in the target URL now restricts the scan to that port only; scans of targets like `https://example.com:9000` will no longer discover or report findings on other ports such as `:81` or `:5013`.

! Single-target scans of a registrable domain no longer enumerate subdomains; subdomain discovery, takeover checks, and sibling subdomain content discovery are silently skipped in single mode.

WAS THIS USEFUL? 👍 👎

## BUILD

◆ AI Coding Agents

# Anysphere Cursor

Feedback ⓘ updated · Nov 7, 2026 NOTES ↗

RANK

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

> Cursor shipped real-time voice chat with coding agents powered by Grok Voice, and launched an Origin App Marketplace for distributing internal and third-party tools like Semgrep and Socket Security to teams.

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

## 01 Origin App Marketplace NEW

81

Cursor added an Origin App Marketplace listing apps across deployments, CI, code review, security, monitoring, and project management, with integrations including Artemis Security, Semgrep, Socket Security, Upwind, Sentry, and Linear. Team admins publish an app via Publish to Marketplace, complete a submission form, and after a security review Cursor deploys it to a chosen URL, which can then be shared with specific people, groups, or the entire team behind Cursor sign-in.

Submit a security or monitoring app to the Origin App Marketplace so your team can install it across deployments and CI pipelines.

> 📍 As a team admin, open the Origin App Marketplace, click Publish to Marketplace, complete the submission form, and await the security review. Once checks pass, Cursor deploys the app to the URL you chose; then share it with specific people, groups, or your entire team, all behind Cursor sign-in.

— Lists integrations and exact publish/review/share workflow. snapshot-20261107

## 02 Real-time voice chat with agents NEW

79

Voice chat is powered by Grok Voice with selectable voices ( Ara , Eve , Leo , Rex , Sal ) configurable in Voice settings; enabling Privacy Mode runs voice chat with zero data retention. During a call, agents speak for themselves — sharing short progress notes while working and a brief summary when done — so the caller hears updates without reading the full written reply. It's available in the IDE, web, and iOS, started via a waveform Start voice chat button next to the microphone button when the chat input is empty.

Start a real-time voice conversation with a coding agent to describe a change out loud and hear progress updates as it works.

📍 In the Agents Window or IDE, with the agent idle and the chat input empty, click the waveform button labeled `Start voice chat` (where the Send button usually appears, next to the microphone/dictation button). Speak your request; the agent will reply out loud and share progress notes. To end the call, click hang up or use the Lock Screen Live Activity on iOS.

Pick a voice and confirm Privacy Mode so voice calls run with zero data retention before using voice chat in a sensitive project.

📍 In Cursor, go to Voice settings and select a voice from Ara, Eve, Leo, Rex, or Sal. Then enable Privacy Mode in settings — with Privacy Mode on, voice chat runs with zero data retention.

— Names mechanism, voices, privacy mode, and exact UI trigger. `snapshot-20261107`

---

WAS THIS USEFUL? 👍 👎

# SST OpenCode

Feedback ⓘ updated · Nov 7, 2026 NOTES ↗ RANK

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

> OpenCode's latest snapshot adds git worktree isolation and archiving for scripted runs, a percentage-based compaction buffer that works across model context sizes, an experimental artifact-publishing mode, shared MCP connection pooling, and expanded GUI plugin capabilities.

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

01 **Percentage-based compaction buffer** NEW 85

Set `"compaction": { "buffer": "20%" }` in `opencode.json` to reserve a percentage of each model's context window, so the same config fits models with both small and large windows without separate token-count settings.

Reserve 20% of every model's context window as a compaction buffer so the same config works across models with very different window sizes.

json

```
{
  "compaction": {
    "buffer": "20%"
  }
}
```

— Named config key with value and clear mechanism. snapshot-20261107

02 **Worktree isolation and archiving for scripted runs** NEW 80

Run scripted sessions in a new git worktree with `opencode run -w` or name it with `--worktree=<name>`; add `--archived` to `opencode run` to keep scripted and CI sessions out of the inbox.

Run a scripted prompt in an isolated git worktree and archive the session so it stays out of your inbox — useful for parallel CI jobs on the same repo.

```
$ opencode run --worktree=feature-branch --archived 'Review the diff and suggest fixes'
```

— Named flags plus a runnable example command. snapshot-20261107

03 **Artifact publishing mode** NEW 75

Enable artifact publishing with `"experimental": { "artifacts": true }` in `opencode.json`, letting agents publish, update, and serve HTML or static artifacts during a session.

Let agents publish and serve HTML artifacts locally during an agentic session, useful for reviewing generated reports or dashboards in the browser.

```json
{
  "experimental": {
    "artifacts": true
  }
}
```

— Named config flag and what it enables. snapshot-20261107

04 **Experimental feature flag controls** NEW 65

Turn on an experimental bytecode VM for Code Mode with `experimental.codemode_vm: true` or `OPENCODE_FLAGS=codemode_vm`, or enable all feature flags at once with `OPENCODE_FLAGS=all` or `experimental.all_flags`.

— Named flags but no example or stated effect. snapshot-20261107

05 **SDK job listing for background tasks** NEW 65

List background shell, subagent, and image jobs via the new SDK method client.experimental.job.list(), with filters for session, descendant, kind, and status.

— Named SDK method and filters, no usage example. snapshot-20261107

THINNER COVERAGE BELOW

06 **Shared MCP connection pooling** IMPROVED 45

Share one MCP server connection across projects with the same setup by turning on pooling, reducing process count and startup time.

— Explains benefit but no config key named. snapshot-20261107

07 **GUI plugin capabilities expanded** IMPROVED 40

GUI plugins now support color themes, code-block rendering and actions, and native settings and menu components.

— Lists additions but no mechanism or API detail. snapshot-20261107

08 **Web-search policy control per session** NEW 35

Added the `websearch.use` policy action to control web-search behavior per session.

— Named policy action but no further mechanism given. snapshot-20261107

WAS THIS USEFUL?  

## yetone magpie

Feedback ⓘ v0.1.3233 CODE ↗ RANK

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

> Magpie's gateway gained a new OpenAI Decisions API endpoint and a native integration with OpenMausBot, alongside smarter per-model quota handling, paginated call history, and a provider priority field for Smart routing.

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

### 01 Decisions API gateway endpoint NEW 95

Adds a `POST /v1/decisions` endpoint to the gateway that proxies OpenAI's Decisions API — supporting predicate, choice, and score question types with a shared input of text and images — to OpenAI, Azure OpenAI ( `/openai/v1/decisions` ), or a chained magpie instance, selecting only providers that can serve the `decision` kind and the `gpt-6-luna` model.

Route an OpenAI Decisions API request through magpie to whichever configured backend can serve it.

```
$ curl -s http://localhost:PORT/v1/decisions \
  -H 'Content-Type: application/json' \
  -d '{"model":"gpt-6-luna","input":[{"role":"user","content":
[{"type":"input_text","text":"Inspect the product in this
photo."},
{"type":"input_image","image_url":"
<base64>"}]}],"questions":
[{"type":"predicate","name":"visible_damage","instructions":"Is
there visible damage?"},
{"type":"choice","name":"condition","instructions":"Its
condition","choices":[{"value":"new"},{"value":"used"}]}]}'
```

— Named endpoint, model, and provider paths with a runnable curl example v0.1.3233

### 02 Native engine registration in OpenMausBot NEW 80

magpie now registers as a native `claudeAgent` engine inside OpenMausBot, writing a `custom -access` instance with `requireApiKey` , `ANTHROPIC_BASE_URL` , `ANTHROPIC_AUTH_TOKEN` , and `ANTHROPIC_API_KEY` set, and all magpie models listed in `endpointModels` — stored in OpenMausBot's `config.json` under `$OMB_DATA_DIR` or `~/.openmausbot` .

— Exact config keys and file path named, but no runnable command v0.1.3221

03 **Lead provider field for Smart routing** NEW 75

Providers now support a `Lead` field that, when set, causes Smart routing to try that provider before all others in its group — useful for exhausting a pay-as-you-go API key's monthly credit before falling back to a subscription.

— Named config field with clear use case, no exact UI steps given v0.1.3208

04 **Paginated Recent Calls list** IMPROVED 60

The gateway UI Recent Calls list now paginates: it defaults to 20 entries, shows a `Show N more` button when additional calls are held in memory, and expands 20 at a time up to 40 total.

— Concrete limits given but only a UI path, not a command v0.1.3213

05 **Per-model quota rest scoping** IMPROVED 60

The gateway now rests only the specific model — not the whole account — when a full quota window covers some models only (pooled buckets), so other models with remaining quota continue routing normally.

— Mechanism explained but it's automatic, nothing for reader to invoke v0.1.3213

WAS THIS USEFUL?  

## keploy

Feedback ⓘ v3.6.224 NOTES ↗ RANK

Open-source platform for creating safe, isolated production sandboxes for API, integration, and E2E testing.

> Keploy refined its mock and replay engine to stop over-consuming mocks on skipped tests and unread responses, and improved how it learns call scope and orders changes during test recording.

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

### 01 Mock and replay handling for skipped tests IMPROVED

40

Keploy's mock system now keeps a skipped test's reads unserved so a wrapping build does not consume mocks belonging to a skipped test, and the replay engine no longer requires a call for an answer the client never read, reducing false failures on unconsumed responses.

— Describes behavior change but no concrete config or command to act on v3.6.224

### 02 Test recording scope and ordering consistency IMPROVED

33

Keploy now learns a call that trails a test and includes it within that test's recording scope, and decides the order of each changed word and call instance in a single consistent pass during learning.

— Explains internal learning behavior with no named surface to act on v3.6.223

WAS THIS USEFUL? 👍 👎

# Letta Code

Feedback ⓘ v0.34.44 NOTES ↗ RANK

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

> Letta Code shipped subagent context isolation by default and expanded tracing to cover external agent model calls, client-side API cost computation, and permission mode transitions.

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

01 **Client-side api_cost_usd computation** IMPROVED 55

Letta Code now computes `api_cost_usd` client-side for both local and cloud agents, making API cost visible without a round-trip to the server.

— Names exact field and mechanism but no access path shown v0.34.42

02 **Subagent context forking by default** IMPROVED 40

Letta Code now forks a new subagent context by default when spawning subagents, isolating each subagent's execution context without requiring explicit configuration.

— Describes behavior change but no config detail or mechanism v0.34.44

03 **Tracing for external agent model calls** NEW 35

Letta Code now captures model calls made by external coding agents in tracing, making third-party agent activity visible alongside native agent traces.

— Clear capability but no interface or access path given v0.34.44

04 **Permission mode transition tracing** NEW 35

Letta Code's listener now traces permission mode transitions, giving practitioners a log of when and how permission modes change during agent execution.

— Describes what is logged but not where to view it v0.34.42

WAS THIS USEFUL? 👍 👎

# opencode-telegram-bot

Feedback ⓘ v0.27.0 NOTES ↗

RANK

opencode-telegram-bot is a Telegram client for OpenCode that lets you run and monitor AI coding tasks from your phone while they run locally.

> opencode-telegram-bot added an autonomous `/goal` command with run controls, search across sessions and projects, recurring task toggling, background session muting, and standalone OpenCode server support.

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

## 01 Mute background session notifications in /settings NEW

65

A new `/settings` option mutes notifications from background sessions while keeping replies, questions, and permission requests from the active session visible.

Silence distracting notifications from sessions running in the background while staying informed about the session you are actively working in.

📍 `/settings`

— Exact command plus usage example for this toggle v0.27.0

## 02 Standalone OpenCode server via /opencode_start NEW

65

`/opencode_start` runs a standalone OpenCode server without replacing the background service already registered, allowing a secondary server for another project.

Start a secondary OpenCode server for a separate project without touching the background service already registered.

📍 `/opencode` _start

— Exact command with example scenario provided v0.27.0

## 03 Autonomous goal execution via /goal NEW

60

The new `/goal` command lets you enter and confirm an objective that runs autonomously, with Pause, Continue, and Delete controls to manage it while it executes.

— Describes behaviour and controls but no walkthrough example v0.27.0

## 04 Toggle recurring tasks in /tasklist IMPROVED

60

`/tasklist` now lets you disable and re-enable recurring tasks without deleting them, with alarm and sleep icons indicating whether each task is enabled or disabled.

— Explains mechanism and icon indicators, no example given v0.27.0

THINNER COVERAGE BELOW

05 **Search in /sessions and /projects** NEW 55

You can now find sessions by title or ID in `/sessions`, with the current session marked in both the list and search results, and find a project by name using the Search option in `/projects`.

— Names both commands but lacks example output v0.27.0

WAS THIS USEFUL? 👍 👎

# coder_eval

Feedback ⓘ v0.12.17 NOTES ↗ RANK

Playwright for coding agents. Test that your skills, MCP servers, and CLIs actually work when an agent uses them — sandboxed YAML suites, A/B experiments, CI gates.

> coder_eval v0.12.17 adds Docker network egress isolation for agent tasks, scripted multi-turn user simulations, and configurable retry limits for whole agent turns.

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

## 01 LLM-only network egress isolation for Docker tasks NEW

86

Set `network: llm_only` and `egress_allowlist` in `DockerDriverConfig` to restrict a task's Docker container to outbound traffic only toward its model-host endpoint (e.g. `api.anthropic.com`), preventing agents from reaching arbitrary hosts during evaluation; a separate allowlist can be configured for the grading phase.

Lock a task's Docker container to only reach its LLM API endpoint, keeping agents from phoning home to arbitrary hosts during evaluation.

yaml

```yaml
network: llm_only
egress_allowlist:
  - api.anthropic.com
```

— Names exact config keys and a runnable YAML example v0.12.17

## 02 Scripted user turns via simulation.turns NEW

72

Define fixed, scripted user messages for a multi-turn dialog by setting `simulation.turns` in a task YAML file, so the simulated user sends pre-written messages instead of invoking an LLM for each turn.

Pre-script the user side of a multi-turn dialog so the simulated user sends fixed messages instead of calling an LLM for each turn.

yaml

```yaml
simulation:
  turns:
    - 'Please implement a function that reverses a string.'
    - 'Now add type annotations.'
```

— Names config key with concrete YAML usage example v0.12.17

THINNER COVERAGE BELOW

03 **Configurable whole-turn agent retry limits** NEW 33

Task configuration now supports setting a retry limit for whole agent turns, allowing control over how many times an agent turn is retried per task.

— No config key or example given, only described in prose v0.12.17

WAS THIS USEFUL? 👍 👎

## forgedock

Feedback ⓘ v1.51.0 NOTES ↗ RANK

ForgeDock — autonomous AI dev pipeline for Claude Code. GitHub as agent memory: issue in, PR out.

> ForgeDock added an early alert for batches that fail within their first hour, surfacing short-lived failures that previously went unnoticed until a manual check.

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

01 **Early batch-failure alert** NEW 28

ForgeDock now alerts on batches that fail within their first hour, so short-lived failures no longer go unnoticed until a manual check.

— Describes behaviour but no mechanism, threshold detail, or config surface. v1.51.0

WAS THIS USEFUL? 👍 👎

## mvschwarz OpenRig

Feedback ⓘ v0.7.9 NOTES ↗ RANK

OpenRig is a network manager for AI coding agents: build persistent teams with roles, shared context and owned work.

> OpenRig v0.7.9 adds per-permit stream recording and new daemon HTTP read endpoints for client authors, turns on stream intake by default, and ships a wide set of CLI correctness fixes, exit-code changes, and upgrade/migration notes spanning export, send, workspace validation, environment commands, Slack reconciliation, and agent guidance delivery.

└──▷ **WHAT SHIPPED · 16 FEATURES** most completely described first ? what's the number?

### 01 Permission ask visibility and human input NEW 85

Adds `rig stream list --tag runner-permission-permit` to surface every permission OpenRig auto-granted without a human present, capturing the tool call, the reason no person answered, and the option that was presented; and forwards text a human types when answering a runner permission ask to the agent as the `said` field in the read result, with a plain yes or no interpreted as allow or reject.

Audit every permission OpenRig granted automatically — without a human present — to review what tools were allowed and why.

```
$ rig stream list --tag runner-permission-permit
```

— Names exact command and tag, runnable as-is Explains mechanism but no command to try it

v0.7.9

### 02 Stream intake now default-on BREAKING 85

Every `rig stream emit` now opens a `queue.worker` intake row by default; run `rig config set stream.intake off` to restore the prior 0.7.8 behaviour.

Suppress the new default stream-intake behaviour so that emitting a stream item does not wake the queue worker, preserving 0.7.8 semantics.

```
$ rig config set stream.intake off
```

— Clear before/after with exact revert command v0.7.9

### 03 Daemon HTTP reads for client authors NEW 85

New daemon HTTP endpoints let client authors read people and the default person, open updates without Slack, a rig's messages in one request, and Slack channels and people without the client holding the bot token; documented in `docs/reference/daemon-http-clients.md`. An opt-in `decisionRevision` field lets a daemon HTTP client pin the exact decision revision its person reviewed. Serve the daemon over HTTP with `rig mcp serve --http --grant-service <url>`.

Expose OpenRig over HTTP so a connected assistant can claim and manage its own queue rows, scoped through your own grant service.

```
$ rig mcp serve --http --grant-service https://your-grant-service.example.com
```

— Names doc path, field and a runnable serve command v0.7.9

04 **Agent guidance delivered outside repo AGENTS.md** BREAKING

85

Fresh Pi seats, and fresh Codex seats on Codex 0.164.0 or later, now receive OpenRig's guidance outside the repository `AGENTS.md`, delivered as launch developer instructions. If only your shell startup sets `CODEX_HOME`, OpenRig reads `~/.codex`, so a `developer_instructions` file in your home is replaced for that session; declare the same absolute path as `CODEX_HOME` in the member's or rig's `env:` to avoid this.

— Detailed mechanism with concrete migration step v0.7.9

05 **Slack connector reconciles ended asks** IMPROVED

70

After upgrading, the Slack connector edits previously ended asks to show they were handled, working within Slack's rate limit of roughly 450 asks per hour; `rig slack status` shows pending edits.

— Names rate limit and a status command to check progress v0.7.9

06 **`rig export` no longer overwrites rig.yaml** BREAKING

65

`rig export` with no `-o` now writes `rig-export.yaml` beside an existing `rig.yaml` instead of overwriting it.

— Clear before/after for a named command v0.7.9

07 **`rig send --verify` reports unconfirmed sends** BREAKING

65

`rig send --verify` now reports unconfirmed and exits 1 (not delivered) when it cannot confirm what its Enter did.

— Exact exit-code behaviour named v0.7.9

---

08 **Ambiguous rig name handling in env commands** BREAKING

65

When an active rig name matches more than one rig, `rig env down`, `status` and `logs` now list the matching rig IDs and stop, rather than silently using the first match.

— Clear before/after across three subcommands v0.7.9

---

09 **OpenShell guest image rebuild required** BREAKING

60

OpenShell environment users must rebuild their guest image using `daemon/assets/openshell/README.md`; an older helper rejects the new startup step.

— Names doc path for the required rebuild v0.7.9

THINNER COVERAGE BELOW

---

10 **Hide built-in specs from views** NEW

55

Set `specs.builtins_visible: false` to hide OpenRig's built-in specs from views.

— Exact config key but no further detail v0.7.9

---

11 **`rig workspace validate` exit code for malformed frontmatter**

55

BREAKING

`rig workspace validate` now exits 1 for malformed optional frontmatter.

— Named command and exit behaviour, little else v0.7.9

---

12 **New default kernel location** IMPROVED

55

New default kernels now start in `<workspace.root>/kernel`; existing kernels stay where they are but can optionally be moved as described in `docs/reference/agent-startup-guide.md`.

— Names path and doc but low urgency for existing kernels v0.7.9

---

13 **409 response for answers to settled asks** BREAKING

55

A later answer to an already-settled ask now returns 409 `qitem_already_terminal` and records nothing.

— Named status code and error key, limited scope detail v0.7.9

14 **Automatic plugin channel forwarding for Claude members** NEW 50

A Claude member's `plugin_channels` are now passed automatically, forwarding them to `--channels`.

— Names field and flag, thin description v0.7.9

15 **`rig down` cleans up OpenRig's additions** IMPROVED 50

`rig down` now removes OpenRig's own additions from a working folder once no live seat uses them.

— Behaviour described but no further mechanism v0.7.9

16 **Nested context install names** NEW 45

`rig context add <url>` can now install a context into a new nested name.

— Runnable command but minimal explanation v0.7.9

## └──▷ BREAKING ON UPGRADE

! Stream intake is now on by default: each `rig stream emit` opens a `queue.worker` intake row. To restore 0.7.8 behaviour run `rig config set stream.intake off` after upgrading.

! The bundled openrig-core plugin advances from 0.1.24 to 0.1.25; managed refresh now preserves replaced files under `$OPENRIG_HOME/backups/managed-refresh/<id>/` — reapply any local edits you still need from that copy.

! OpenShell environment users must rebuild their guest image using `daemon/assets/openshell/README.md`; an older helper rejects the new startup step.

! Four database migrations (146 to 149) run on first start after upgrading; restart the daemon with `rig daemon stop && rig daemon start`.

! `rig export` with no `-o` now writes `rig-export.yaml` beside an existing `rig.yaml` instead of overwriting it.

! `rig send --verify` now reports unconfirmed and exits 1 (not delivered) when it cannot confirm what its Enter did.

! `rig workspace validate` now exits 1 for malformed optional frontmatter.

! An active rig name matching more than one rig now makes `rig env down`, `status` and `logs` list the matching rig IDs and stop, rather than using the first match.

! Fresh Codex seats on Codex 0.164.0 or later receive OpenRig's guidance as launch developer instructions outside the repository `AGENTS.md`; if only your shell startup sets `CODEX_HOME`, OpenRig reads `~/.codex`, so a `developer_instructions` in your home is replaced for that session — declare the same absolute path as `CODEX_HOME` in the member's or rig's `env:` to avoid this.

! New default kernels start in `<workspace.root>/kernel`; existing kernels stay where they are but can optionally be moved as described in `docs/reference/agent-startup-guide.md`.

! `rig down` now removes OpenRig's own additions from a working folder once no live seat uses them.

! After upgrading, the Slack connector edits previously ended asks to show they were handled (at Slack's rate limit of roughly 450 asks per hour); `rig slack status` shows pending edits.

! A later answer to an already-settled ask now returns 409 `qitem_already_terminal` and records nothing.

---

WAS THIS USEFUL? 👍 👎

# Anthropic Claude Code

Feedback ⓘ v2.1.319 NOTES ↗

CODE ↗ RANK

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

> Claude Code's single release this window adds a permission-deny mode, background MCP task handling, a new latency-optimization command, and a cluster of Claude apps gateway admin controls, alongside a breaking change to auto mode's fallback behaviour behind LLM gateways.

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

## 01 Auto mode server-side classifier behind gateways

BREAKING

90

Auto mode behind an LLM gateway no longer falls back to the built-in classifier when telemetry or feature flags are off; tool calls the server cannot check are now asked about or denied instead. Set `CLAUDE_CODE_AUTO_MODE_SERVER=0` to restore the previous client-side classifier behaviour, or `CLAUDE_CODE_AUTO_MODE_SERVER=1` to force server-side checks even when `CLAUDE_CODE_DISABLE_EXPERIMENTAL_BETAS` is set; Bedrock, Vertex and Foundry are unchanged.

Pin Claude Code to built-in feature-flag defaults in a locked-down CI environment without disabling telemetry.

```
$ export CLAUDE_CODE_DISABLE_FEATURE_FLAGS=1
claude -p 'run the test suite and summarise failures'
```

— Names exact env vars, defaults and unaffected platforms v2.1.319

## 02 Deny instead of prompting on permission checks

NEW

85

The `permissions.denyInsteadOfPrompting` setting denies a tool call that would otherwise block on a permission prompt and lets the turn continue, in terminal sessions, subagents and teammates.

Prevent permission prompts from blocking automated or subagent turns by denying the call and continuing instead of waiting.

```json
{ "permissions": { "denyInsteadOfPrompting": true } }
```

— Exact config key with a runnable JSON example v2.1.319

03 **Claude apps gateway admin controls** NEW 85

On the Claude apps gateway, operators can set `CLAUDE_GATEWAY_LOG_FORMAT=json` to emit operational log lines as JSON with `severity`, `time` and `message` fields; control pooled group budget priority with `admin.pooled_conflict_rule` and `admin.pooled_priority_order` to set which pooled group budget pays first for a developer in several groups; and configure a `login_notice` setting to show an operator-defined message on the gateway's browser verification page and on Claude Code's sign-in screen.

Emit Claude apps gateway operational logs as structured JSON so your log collector can parse severity from a field.

```
$ CLAUDE_GATEWAY_LOG_FORMAT=json ./gateway
```

— Names all three distinct config surfaces verbatim v2.1.319

04 **Background MCP tasks in the terminal** NEW 80

When an MCP server runs a tool call as a task, Claude Code now moves it to the background immediately and reports the result to Claude when it finishes. Set `CLAUDE_CODE_DISABLE_MCP_TASK_BACKGROUND` to instead force Claude Code to wait for the MCP task inside its tool call rather than backgrounding it.

— Names the env var that reverses the behaviour v2.1.319

05 **Usage spend view behind a gateway** NEW 60

View spend for today, this week and this month in `/usage` for developers behind a Claude apps gateway who have no spend limit, when the gateway runs this version or later.

— Names command and conditions but no further mechanism v2.1.319

THINNER COVERAGE BELOW

06 **Latency-optimize command for Claude API code** NEW 55

Run `/claude-api latency-optimize` to find where Claude API code introduces latency and try faster settings against your eval.

— Runnable command but no detail on mechanism v2.1.319

07 **Reopen finished subagents in /tasks** IMPROVED 50

An Earlier in this session row in `/tasks` now lists finished subagents so you can reopen one and message it even after it leaves the list or you restart.

— Clear UI path but no deeper mechanism described v2.1.319

└──▷ **BREAKING ON UPGRADE**

! Auto mode with telemetry or feature flags off (including behind LLM gateways) no longer falls back to the built-in classifier; calls the server cannot check are asked about or denied. Set `CLAUDE_CODE_AUTO_MODE_SERVER=0` to restore the previous behaviour. Bedrock, Vertex and Foundry are unchanged.

! Auto mode behind an LLM gateway: if managed settings set `CLAUDE_CODE_DISABLE_EXPERIMENTAL_BETAS`, they must also set `CLAUDE_CODE_AUTO_MODE_SERVER=1` to keep the server-side check enabled.

WAS THIS USEFUL?  

# Alibaba Qwen Code

Feedback ⓘ v0.26.9 NOTES ↗ RANK

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

> Qwen Code now lets subscribers authenticate directly with an existing Qwen Plus or Pro subscription, removing the need for a separate API key.

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

01 **Login via Qwen subscription account** NEW 55

Users can authenticate using an existing Qwen Plus or Pro subscription from `chat.qwen.ai` by selecting Qwen Account with Subscription in Qwen Code, with no API key required.

— Names exact UI option and source, but no further mechanism detail. product docs

NEW SUPPORTED PLATFORM Image input for Llama, LLaVA, Pixtral, Mistral Small models

NEW SUPPORTED PLATFORM Image input for Qwen VL models under Ollama names

NEW INTEGRATION OCI Generative AI documented under OpenAI-compatible providers

v0.26.9 NOTES ↗ LATEST STABLE

Qwen Code adds bulk Skill activation via `PUT /skills/activation`, daemon ZIP install APIs, `--event-ring-max-bytes` memory cap, OSC 7501 terminal status, and `experimental.unifiedAgentLoop` for Ink TUI tool batches.

└──▷ **GET THIS VERSION** ▾

Docker

```
$ docker pull ghcr.io/qwenlm/qwen-code:0.26.9
```

npm

```
$ npm install -g @qwen-code/qwen-code@0.26.9
```

git

```
$ git clone --branch v0.26.9 https://github.com/QwenLM/qwen-code.git
# already have the repo? check out this version:
$ git checkout v0.26.9
```

published under the project's GitHub account · signed build provenance · checked 7 Nov

Wrong command?

Activate or deactivate 1–100 Skills in one call via the new `PUT /skills/activation` endpoint and `setUserSkillsActivation` TypeScript SDK method, with saved preferences and propagation status returned.

Install large Skill ZIP archives via new daemon API routes — by upload or host file path — and query installation receipts after failures.

Install Skill ZIPs from File, Blob, bytes, or host path using the new TypeScript SDK, with attempt IDs and receipt lookup.

Cap daemon command and skill replay snapshot memory with an 8 MiB per-session ring buffer, configurable via `--event-ring-max-bytes`.

Enable tool-batch completion for the Ink TUI by turning on `experimental.unifiedAgentLoop`, so tool continuations finish within the same turn.

+8 more

---

WAS THIS USEFUL? 👍 👎

## flow-next

Feedback ⓘ flow-next-v9.8.0 NOTES ↗ RANK

flow-next is a workflow plugin that takes a bug, idea or ticket to a verified pull request with specs, cross-model review, live QA, and receipts

> flow-next's biggest change this window is a breaking default: specs now go straight to execution instead of automatic task planning. It also added circular-dependency detection to its flowctl spec commands and introduced GitLab merge-request lifecycle commands.

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

01 **Default no-plan execution mode for specs** BREAKING 95

Specs now route directly to execution unless a plan or task breakdown is explicitly requested in the `/flow-next:flow` request or spec body, or by running `/flow-next:plan`. `/flow-next:work` now lists the direct route first as the recommended option when it meets a spec with no tasks and no named route. Skills that previously pointed at `/flow-next:plan` — task-manager, prime, prospect, strategy, and map — now point at `/flow-next:flow`, which picks the route. `flowctl judge --preset route` now returns `work_no_plan_default` for a ready spec with no tasks and no recorded route, and no longer reads the spec body for plan signals. As of 9.8.0, multi-repo, multi-PR, and multi-team specs no longer trigger automatic task planning: a spec that named separate owners or several PRs — which 9.7.0 automatically split into tasks — now runs as a single task. Use `flow-next:work --no-plan` to deliver a feature as a single PR without triggering planning, even on a multi-repo spec, or force a breakdown with `/flow-next:flow <spec-id> please give me a task breakdown` or `/flow-next:plan <spec-id>`.

Deliver a feature as a single PR without triggering task planning, even on a multi-repo spec.

```
$ flow-next:work --no-plan
```

Force planning for a multi-team spec that you still want broken into tasks after upgrading to 9.8.0.

```
$ /flow-next:flow <spec-id> please give me a task breakdown
```

Trigger planning explicitly using the dedicated plan command when a spec names separate owners or multiple PRs.

```
$ /flow-next:plan <spec-id>
```

— Names exact commands, skills, flags and before/after behavior. `flow-next-v9.8.0`

02 **Circular dependency detection in flowctl** NEW 95

`flowctl spec add-dep` now refuses a dependency edge that would close a cycle — including self-dependencies — writes nothing, and names the full cycle path (e.g. `fn-3 -> fn-1 -> fn-2 -> fn-3`) with error code `dependency_cycle`. `flowctl validate --all` lists all cycles already on disk and fails the run, reporting them under `spec_cycles`. Cycle detection also runs during capture before writing, blocking a proposed epic split before any spec is created if its dependency edges would form a cycle.

Catch a cycle before wiring it in — `flowctl spec add-dep` now refuses and prints the full loop path so you can see exactly which specs form the circle.

```
$ flowctl spec add-dep fn-3 fn-1
```

Audit an existing repo for dependency cycles already on disk — useful after importing specs from an external tracker.

```
$ flowctl validate --all
```

— Exact commands, error code and field names with runnable examples. `flow-next-v9.7.0`

└──▷ **BREAKING ON UPGRADE**

! As of 9.8.0, multi-repo, multi-PR, and multi-team specs no longer trigger automatic task planning; planning only runs when explicitly requested in the request or spec.

! A spec that named separate owners or several PRs — which 9.7.0 automatically split into tasks — now runs as one task. To restore the breakdown, explicitly ask for a plan or task breakdown in your `/flow-next:flow` request or spec, or run `/flow-next:plan`.

WAS THIS USEFUL? 👍 👎

# Yuval Sarel cones

Feedback ⓘ v0.4.11 NOTES ↗ RANK

Cones is a terminal dashboard for managing coding-agent sessions across projects.

> Cones improved its session-preview experience, adding a live styled follow pane for sessions running in their own terminal and clearer labeling for unattachable or image-paste sessions.

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

## 01 Live styled preview for own-terminal sessions IMPROVED

55

Pressing Enter on a session running in its own terminal now opens the same styled read-only preview used by history rows, following the conversation live instead of showing a plain-text pane. If no readable conversation exists, the UI now shows `own terminal` instead of offering a `watch` option.

— Describes UI behaviour change with named label but no deeper mechanism v0.4.11

## 02 Image-paste session title labeling IMPROVED

40

Session titles now show `[image]` instead of a temp file path when a session opens with a pasted clipboard image.

— Small, clearly named UI label change with limited scope v0.4.10

WAS THIS USEFUL? 👍 👎

# codealta

Feedback ⓘ 4.0.0 NOTES ↗ RANK

Agentic Development Environment

> CodeAlta's 4.0.0 release introduces Fleet for multi-machine agent sessions, a redesigned Adaptive multi-model critic system, broader automation triggers, and a headless CLI, alongside mobile push notifications and desktop session search/archiving added across the two releases in this window.

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

## 01 Fleet management for multi-machine agent sessions NEW

95

Run `altanode run` to start CodeAlta Node ( `altanode` ) on a headless or always-on machine and manage it from Fleet in Settings > Fleet > Machines, with no account or server required for direct pairing. Run `altaserver run` to start CodeAlta Server ( `altaserver` ) on a machine you control, giving teams shared machines, secrets, budgets and audit logs with end-to-end encrypted session relay. Use `alta fleet` to let agents pick machines, start sessions remotely, move sessions and read fleet health, within the limits each machine accepts.

Put an always-on build machine into your fleet so agents can delegate iOS or GPU workloads to it without you being at your desk.

```
$ dotnet tool install -g CodeAlta.Node
altanode run
```

Give a whole team shared machines, budgets and an audit log — install the server on a machine you control and let CodeAlta Desktop connect to it.

```
$ dotnet tool install -g CodeAlta.Server
altaserver run
```

— Names exact binaries, commands and UI path with mechanism and runnable examples 4.0.0

## 02 Adaptive multi-model critic reviews BREAKING

85

Configure Adaptive multi-model reviews in Settings > Models > Adaptive with Low, Medium, High and Max levels, each pairing a working model with read-only critics — including critics from a different provider — that inspect diffs and run tests after every code-changing turn. This replaces the previous Adaptive tier-escalation levels, so existing Adaptive setups must be reconfigured.

.

— Clear mechanism and exact UI path, but no CLI/config key given 4.0.0

03 **CodeAlta CLI for headless agent runs** NEW 75

Adds `altacli` (CodeAlta CLI) for running agents without a window, for scripts, CI and editor integrations.

Run a coding agent from CI or a script without opening any window.

```
$ dotnet tool install -g CodeAlta.Cli
altacli run "Fix the failing test"
```

— Named binary with runnable install and usage command 4.0.0

04 **Desktop session search and archiving** NEW 75

On desktop, find within a session with Ctrl+F or the `/find` command, with every match marked and counted, search across all sessions with Ctrl+Shift+F, and archive the selected session with Ctrl+Shift+A or the `/archive` command.

— Exact keybindings and command given, with match-count behavior Exact shortcut and command but no further mechanism described

4.0.0

05 **Expanded automation triggers and Autonomy mode** NEW 70

Start automations from Jira queries, Slack mentions, incidents or webhooks (not only clock or git events), with an Autonomy mode that caps tool calls, sets approval timeouts, and supports dry-run recording via Try it (dry run).

.

— Names triggers and Autonomy controls but no exact config path 4.0.0

THINNER COVERAGE BELOW

06 **Mobile push notifications and Dev Tunnels support** NEW 50

The mobile app now adds push notifications when a session finishes or needs input, plus support for Dev Tunnels as an alternative to Tailscale for reaching your desktop.

.

— Describes what changed but no setup steps given 4.0.0

07 **Claude Code subagents as prompts** NEW 50

Claude Code `.claude/agents` subagents are now available as agent prompts.

— Names the exact file path but no usage steps 4.0.0

08 **Secrets storage for configuration** NEW 45

Store credentials in Settings > Secrets and reference them in configuration instead of keeping tokens in files.

– UI path given but no config key or reference syntax 4.0.0

09 **Inspectable terminal paste** NEW 40

Paste in terminal puts a shell block into the terminal for inspection without executing it automatically.

– Behavior described but no command or flag named 4.0.0

10 **Hide models from a provider** NEW 35

Hide models from a provider using rules, with the ability to keep exceptions visible.

– States capability without naming the rule mechanism 4.0.0

└──▷ **BREAKING ON UPGRADE**

! Adaptive tier-escalation levels are replaced by the new Low/Medium/High/Max critic levels; existing Adaptive setups must be reconfigured in Settings > Models > Adaptive.

WAS THIS USEFUL? 👍 👎

## All Hands AI OpenHands

Feedback ⓘ v1.37.0 NOTES ↗

RANK

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

> OpenHands added Mermaid diagram visualizations for automations in this release.

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

01 **Mermaid diagrams for automation flows** NEW 35

OpenHands now generates a Mermaid diagram for each automation to visualize its flow at a glance, shown in the UI.

– Names the feature but no detail on mechanism or location v1.37.0

WAS THIS USEFUL? 👍 👎

# pstack-claude

Feedback ⓘ v0.9.131 NOTES ↗ RANK

pstack-claude is a port of Poteto's pstack agent framework to Claude Code, Codex, Copilot, Pi, and other AI coding assistants.

> pstack-claude's latest release adds subagent support for Grok Build, tightens an internal test to catch stale model-panel mappings, and retires a panel role that upstream removed.

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

01 **Grok Build subagent support and messaging** NEW 92

Adds `spawn_subagent` and `send_subagent_message` commands for Grok Build. `spawn_subagent` takes a `run_in_background` option defaulting to true, across three updated rows in `grok.md`. `send_subagent_message` is enabled for Grok via the `GROK_ACTIVE_AGENT_MESSAGES` environment variable or by adding `active_agent_messages` under `[features]` in config. Dispatching to `general-purpose` or `pstack:poteto-agent` now omits `subagent_type` in `grok.md`, since Grok runs an omitted type as general-purpose and cannot select it by name.

– Names flags, env var and config key but no full run example v0.9.131

02 **Retired `how critics` panel role** BREAKING 60

The `how critics` panel role has been retired, synced from upstream commit e8d856f (PR #67); any mapping or config that referenced `how critics` as a panel-tier skill will no longer match.

– Names the removed role and source commit but no migration path given v0.9.131

THINNER COVERAGE BELOW

03 **Panel-skill mapping test for runtimes** IMPROVED 59

`tests/models.test.mjs` now compares each runtime mapping's panel list against skills that carry a `panel` role in `models.json`, so a stale mapping fails the test automatically.

– Describes mechanism but is an internal test, not user-facing action v0.9.131

└──▷ **BREAKING ON UPGRADE**

! The `how critics` panel role has been retired (synced from upstream e8d856f, PR #67); any mapping or config that referenced `how critics` as a panel-tier skill will no longer match.

WAS THIS USEFUL? 👍 👎

## repoprompt-ce

Feedback ⓘ v1.9.0 NOTES ↗ RANK

Repoprompt-ce is a native macOS context engineering app for AI coding agents, with an MCP CLI (community edition of RepoPrompt).

> RepoPrompt CE v1.9.0 reworks the Agent Mode sidebar with dedicated Chats and Overseers tabs and nested session visibility, while also surfacing independently-started Claude Code turns and smoothing session-activation focus behavior.

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

01 **Redesigned Agent Mode sidebar** IMPROVED 53

The Agent Mode sidebar now has separate Chats and Overseers tabs, nested overseen sessions, improved header controls, and direct navigation from Overseer badges.

— Names UI tabs and controls but no mechanism detail v1.9.0

02 **Session activation focus handling** IMPROVED 50

Session activation now focuses pending questions or the composer while preserving live typing and keeping Return away from pending approvals.

— Describes behavior and edge cases but no concrete surface v1.9.0

03 **Independent Claude Code turn visibility** IMPROVED 35

Claude Code turns that start independently are now visible and controllable as ordinary agent runs.

— States behavior change but lacks mechanism or UI path v1.9.0

WAS THIS USEFUL? 👍 👎

## Sourcegraph Amp

Feedback ⓘ Close the Loop (2026-11-06)

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 shipped two new collaboration surfaces: an in-page feedback widget that turns any webpage into an Amp thread starter, and Spaces, a multi-agent layer where one coordinating agent manages threads for a team.

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

01 **In-page feedback widget for websites** NEW 70

Appending `?amp` to a site URL activates an annotation widget that lets any visitor click an element, write instructions, and send feedback that automatically opens a new Amp thread in the project. Visitors connect their Amp account once via a Continue with Amp panel before choosing where Amp works and sending.




— Names exact activation method but no API/config details. Close the Loop (2026-11-06)

02 **Spaces multi-agent coordination layer** NEW 65

Spaces is a conversation layer where one coordinating agent spawns and manages sub-agent threads, removing the need to create and organize threads manually. Spaces support team collaboration — multiple users can drop feedback, bugs, and ideas into a shared space, with the agent delegating items into threads and members chatting via comments. A space can be given a standing goal (e.g. monitoring user issues, system health, or security reviews), and the agent splits work across threads while maintaining a summary artifact for review.




— Explains mechanism and use cases but no named config or command. Spaces (2026-11-06)

---

WAS THIS USEFUL? 👍 👎

## Amazon Kiro

Feedback ⓘ

CLI: Compaction Hooks and Session Credits (2026-11-05) BLOG ↗ RANK

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

> Kiro shipped Compaction Hooks across both the CLI and IDE to intercept conversation summarization, persistent cross-session Memory that sub-agents can now read, and new Workflow Run tooling (a run tab with step tree/transcript and a search/filter field), alongside session credit tracking, tighter MCP credential permissions, and agent file-resource limits.

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

### 01 Persistent Memory across chat sessions and sub-agents NEW

91

Kiro now carries Memory across chat sessions, controlled via the `Kiro Agent: Memory` setting, including a `Use only` mode that lets Kiro read memories without changing them. A new Memory view, opened from the Memories icon in the activity bar, shows what Kiro remembers, and chat indicates when Kiro accesses or changes a memory. Sub-agents can now read the memories of the agent that dispatched them, read-only — they cannot add, change, or delete parent memories.

Prevent Kiro from writing new memories while still letting it apply existing ones — useful in shared or regulated environments.

> 📍 In the IDE, open Settings and set `Kiro Agent: Memory` to `Use only`.

Prevent Kiro from writing new memories mid-session while still letting it draw on existing ones — useful when you want stable, read-only context during a focused task.

> 📍 In the IDE, open Settings and set `Kiro Agent: Memory` to `Use only`, then open a chat session — Kiro will read existing memories but will not add or update them.

— Names the setting, its mode, and the new UI view.

IDE 1.4.121: Memory Across Chat Session… | CLI: Compaction Hooks and Session Credi… | IDE: Compaction Hooks, Workflow Run Sea… | IDE 1.4.86: Memory Across Chat Sessions…

### 02 Compaction Hooks for conversation summarization NEW

90

Kiro adds `PreCompact` and `PostCompact` Hooks in the CLI, and `Pre Compaction` / `Post Compaction` triggers in the `Compaction Hooks` group of the IDE Hook

editor. Hooks can inspect the conversation and refuse compaction: a Pre Compaction Hook runs before the summary is written and can cancel compaction by exiting with code 2, while a Post Compaction Hook runs after the summary replaces older history, with its output added to the conversation so the agent retains what the summary left out.

Prevent a compaction from running automatically — for example, to block it during a critical build — by exiting with code 2 from a Pre Compaction Hook.

> 📍 In the Kiro Hook editor, open the Compaction Hooks group and add a new hook with the `Pre Compaction` trigger. In the hook script, write your guard logic and exit with code 2 (and an error message to stderr) to cancel the compaction.

Append context that the compaction summary may drop — such as open TODOs or decision logs — back into the conversation automatically after each compaction.

> 📍 In the Kiro Hook editor, open the Compaction Hooks group and add a new hook with the `Post Compaction` trigger. Write your script to output the content you want preserved; Kiro adds that output to the conversation after the summary is written.

— Exact trigger names and the exit-code-2 cancellation mechanism are given.

IDE: Compaction Hooks, Workflow Run Sea… | CLI: Compaction Hooks and Session Credi…

---

03 **System-wide default settings via KIRO_DEFAULT_SETTINGS** NEW

81

Set `KIRO_DEFAULT_SETTINGS` to the absolute path of a root-owned settings file on macOS or Linux to ship default configuration as the lowest-priority layer, overridden by each user's own `cli.json`.

Lock down a shared dev machine so every Kiro user inherits a common baseline config (e.g. approved model settings) without being able to accidentally override it at the system level.

```
$ export KIRO_DEFAULT_SETTINGS=/etc/kiro/defaults.json
```

— Names the exact env var and override precedence.

CLI: Compaction Hooks and Session Credi…

---

04 **Workflow Runs: search/filter and run detail tab** NEW

75

Workflow Runs gains a `Search workflows` field above the run list to filter by name, status, or recorded duration (cleared via `Clear search` or Escape), plus a new Workflow tab opened via `Open run` on a run's card, showing a step tree on the left, the selected step's transcript on the right, and Pause, Resume, or Stop controls in the tab header.

Review every step of an in-progress workflow run alongside its transcript without leaving context.

> 📍 1. Enable Workflows in Kiro settings. 2. In chat, locate a run's card and select Open run. 3. The Workflow tab opens: the step tree appears on the left, the selected step's transcript on the right. 4. Use Pause, Resume, or Stop in the tab header to control the run.

– Names the exact field and clear mechanism. Exact navigation and controls are walked through.

IDE: Compaction Hooks, Workflow Run Sea… | IDE: Workflow Run Tab and Steering Warn…

---

05 **Warnings for problematic steering documents** NEW 70

The Agent Steering view now flags steering documents Kiro could not fully apply with a warning icon and a problem count; hovering over the warning shows each problem description and a suggested fix.

Identify and fix steering documents that Kiro could not fully apply before they silently affect agent behaviour.

> 📍 1. Open the Agent Steering view in the Kiro IDE sidebar. 2. Look for any steering document marked with a warning icon and a problem count. 3. Hover over the warning to read each problem description and the suggested fix. 4. Edit the steering document to resolve the flagged issues.

– Exact UI location and inspection steps given. IDE: Workflow Run Tab and Steering Warn…

---

06 **File resource limits for agent context** IMPROVED 63

Agent `file://` resources are now capped at 32 files and 100,000 characters, with large files cut or skipped.

– Concrete limits given but no configuration or override shown.

CLI: Compaction Hooks and Session Credi…

---

07 **Session recap after idle time** NEW 60

Enable Session recap under `/settings > Features` or set `chat.enableSessionRecap` to see a short recap above the prompt after idle seconds.

Turn on the session recap so a summary appears above the prompt after you step away, without manually running `/usage`.

> 📍 In the Kiro CLI, run `/settings` › Features and enable Session recap.

– Names the setting and toggle path but not the idle threshold.

CLI: Compaction Hooks and Session Credi…

---

08 **Per-session credit tracking** NEW 60

Track session credit consumption in `/usage`, in the status line via the Session credits segment under `/settings display`, and per step in the workflow monitor.

— Names three surfaces but no example of use. CLI: Compaction Hooks and Session Credi…

09 **Owner-only permissions for MCP credentials** IMPROVED 60

On macOS and Linux, remote MCP registration and credential files are now readable and writable only by the owning user, and new credential folders are created with owner-only permissions.

— Clear security change but nothing for a reader to action.

CLI: Compaction Hooks and Session Credi…

THINNER COVERAGE BELOW

10 **Markdown-first web fetching** IMPROVED 50

`web_fetch` now requests Markdown (or HTML) from sites, preserving code blocks and links in documentation pages.

— Behavior described but no flag or example to invoke it.

CLI: Compaction Hooks and Session Credi… IDE: Compaction Hooks, Workflow Run Sea…

11 **Consolidated notification queue** IMPROVED 45

Notifications now queue into a single row, most urgent first, with a `+N` waiting count; warnings and errors stay on screen until dismissed.

— UI behavior described without a config toggle. CLI: Compaction Hooks and Session Credi…

12 **Kiro Agent version in About dialog** IMPROVED 25

The About dialog now lists the Kiro Agent version below the build date.

— Bare cosmetic change with no further detail. IDE: Workflow Run Tab and Steering Warn…

WAS THIS USEFUL?  

◆ Coding Agent Add-ons

# JuliusBrussee Caveman

Feedback ⓘ bin-v4.1.0 NOTES ↗

RANK

why use many token when few token do trick. Viral skill + proxy for coding agents that cuts 65% of tokens by talking like a caveman.

> Caveman expanded its code compressor to six languages beyond Go and shipped signed cross-platform runtime binaries under Apache 2.0.

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

01 **Code compression for more languages** IMPROVED 60

The code compressor in the `caveman-proxy` and `caveman-engine` binaries now handles TypeScript, JavaScript, Python, Rust, Java, and C/C++, in addition to the existing Go support.

— Names binaries and languages but no mechanism detail. bin-v4.1.0

THINNER COVERAGE BELOW

02 **Signed binaries for macOS, Linux, Windows** NEW 45

Signed runtime binaries for macOS, Linux, and Windows are now available under the Apache License 2.0.

— States platforms and license but no download path or usage. bin-v4.1.0

WAS THIS USEFUL? 👍 👎

## sirmalloc ccstatusline

Feedback ⓘ v2.2.34 NOTES ↗

RANK

A customizable statusline for Claude Code CLI with powerline support, themes, and widget-based configuration.

> ccstatusline's v2.2.34 release adds a context-window override environment variable, several new usage/cache widgets, a thin context-bar mode, Zed IDE link support, and a round of widget/TUI customization improvements, while removing manual terminal-width configuration in favor of reading it from Claude Code directly.

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

### 01 CCSTATUSLINE_CONTEXT_WINDOW_SIZE environment override NEW

80

Introduces the `CCSTATUSLINE_CONTEXT_WINDOW_SIZE` environment variable to override the context window size used in ratio calculations, for cases where Claude Code reports an unexpected model context limit.

Override the context window size when Claude Code reports an unexpected model context limit.

```
$ export CCSTATUSLINE_CONTEXT_WINDOW_SIZE=200000
```

— Named env var with a runnable export command. v2.2.34

### 02 Widget and TUI customization improvements IMPROVED

73

Adds a short time toggle and drops clock notation in the cache timer widget, lets you pick 256-color colors from a grid in the Edit Colors TUI screen, adds optional brackets and per-level colors to the thinking-effort widget, and extends speed widget windows up to one hour (previously capped below an hour).

— Names concrete widgets and the TUI screen affected. v2.2.34

THINNER COVERAGE BELOW

03 **New cache-miss and usage-limit widgets** NEW 55

Adds a prompt cache miss widget to track cache misses alongside existing cache metrics, plus Weekly Limit Timer and Weekly Fable Limit Timer widgets in the usage section, and an Extra Usage Today widget to surface daily extra usage at a glance.

— Widgets named but no setup steps given. v2.2.34

04 **Terminal width now read from Claude Code's COLUMNS** BREAKING 55

Terminal width is now read from Claude Code's `COLUMNS` value, and the Terminal Width options have been removed.

— Clear before/after but no migration command. v2.2.34

05 **Zed IDE link support** NEW 45

Adds Zed as an IDE link mode so the current directory opens in Zed.

— Names the mode but not the exact setting location. v2.2.34

06 **Thin line context bar display mode** NEW 35

Adds a thin line display mode option to the context bar.

— Bare description, no config path given. v2.2.34

└──▷ **BREAKING ON UPGRADE**

! The Terminal Width options have been removed; terminal width is now taken from Claude Code's `COLUMNS` value.

WAS THIS USEFUL? 👍 👎

## Diagram Design

Feedback ⓘ changes since 2026-10-13 CODE ↗

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

Diagram Design adds architecture containment scopes and native Qoder plugin support

WAS THIS USEFUL? 👍 👎

◆ Model Training & Fine-Tuning

# fizgig

Feedback ⓘ v9.0.8 NOTES ↗ RANK

Fizgig — LoRA & Fine-tune Studio for Klein 9B, Krea 2, MiniMax H3, Qwen Image 2.1, Anima & Z-Image Turbo and Ace-Step XL: train, fine-tune, profile, repair and extract LoRAs & LoKR for Image, Video & Music Models

> Fizgig v9.0.8 makes the Muon optimizer the default for every built-in LoRA, slider, and edit preset across eleven image models, while also adjusting Fibo 2.0's training learning rate.

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

01 **Muon optimizer default for LoRA, slider and edit presets**

NEW SUPPORTED PLATFORM IMPROVED 80

Fizgig now defaults every LoRA, slider, and edit preset to the Muon optimizer for Klein 9B, Flux 3 Image Klein, Fibo 2.0, Krea 2, Qwen Image 2.1, MOSS-Image, Z-Image Turbo and Base, SDXL, Anima, and Vimoo 1.0 Lite; reload any preset to apply it, while previously saved presets and the Training form keep the optimizer they had. Muon uses less memory than `AdamW` and slightly more than `AdamW8bit` (about 0.2 GB on a rank-32 Krea 2 LoRA) by storing one value per trained parameter instead of two; `AdamW8bit` and `AdamW` remain available in the Optimizer Type row.

— Names models, memory metric and reload step but no command v9.0.8

02 **Fibo 2.0 learning rate change** IMPROVED 60

Fibo 2.0's main preset now trains at a flat learning rate of `1e-4` instead of Automagic's `1e-6` start; Automagic remains selectable in the Optimizer Type row.

— Gives exact before/after values but limited to one preset v9.0.8

v9.0.7 NOTES ↗ STABLE

Fizgig adds Vimoo 1.0 Lite training, batch size Auto on 40 GB+ cards for five models, and a Studio deferred-render Update button.

└──▷ GET THIS VERSION ▾

Docker

```
$ docker pull ghcr.io/shootthesound/fizgig:9.0.7
```

```git
$ git clone --branch v9.0.7
https://github.com/shootthesound/Fizgig.git
# already have the repo? check out this version:
$ git checkout v9.0.7
```

published under the project's GitHub account · checked 7 Nov Wrong command?

Batch size now has an Auto option for SDXL, MiniMax H3, Anima, Flux 3 Image Klein and Fibo 2.0 on 40 GB+ cards, picking the measured fastest batch per step and reporting the choice under the batch size box.

Gradient checkpointing now turns off automatically on 40 GB+ cards for SDXL, Anima, Flux 3 Image Klein and Klein when the run fits, running 1.3–2.3× faster.

Studio now holds renders until you click Update, letting you change size, clip length, steps, seed, reference picture, negative prompt and LoRA strengths together before triggering a render.

Gizmo now asks which video model a clip is for and prepares it to that model's frame counts and sizes.

The Turbo Preview option in Studio is now called Reuse unchanged blocks and moves its stored results to system RAM when the card is tight, so long clips no longer run out of memory with it enabled.

---

WAS THIS USEFUL? 👍 👎

# NVIDIA NeMo RL

Feedback ⓘ v0.8.0 NOTES ↗ RANK

NeMo RL provides distributed reinforcement-learning and preference-optimization workflows for training language models.

> NeMo RL's v0.8.0 release introduces a Single Controller async RL runtime with streaming training, pluggable samplers and fault tolerance, new TensorRT-LLM and Dynamo rollout engines, and faster NCCL/NIXL weight-refit transports, alongside OpenTelemetry observability, NVFP4/W4A4 quantized training on Blackwell, LoRA fine-tuning, and new audio-visual and Nemotron 3 Ultra training recipes.

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

01 **Single Controller async RL runtime** NEW 100

NeMo RL adds a Single Controller async RL runtime for GRPO or PPO that overlaps rollout and training using a `TransferQueue` data plane; set `async_rl.min_groups_for_streaming_train` below `num_prompts_per_step` to begin training once that many prompt groups finish. Choose which rollout groups the trainer samples via `async_rl.sampler.name` ( `in_order` , `weight_fifo` , `ready_first` , `windowed` , or `custom` ), where `ready_first` reuses late groups instead of discarding them. Configure fault tolerance through `async_rl.rollout_failure` , `async_rl.generation_fleet_health` , and `async_rl.generation_router` to retry or drop failed rollouts and rebuild generation shards on the fly; checkpoint and resume including in-flight rollouts with periodic snapshots on `simple` or `mooncake_cpu` data-plane backends; and capture exact tokens and logprobs (including images and video on vLLM) from NeMo Gym rollouts without retokenization. Run it with `uv run examples/run_grpo_single_controller.py --config examples/configs/grpo_math_1B_megatron_single_controller.yaml` , requiring `data_plane.enabled: true` and, for vLLM, `async_engine: true` with non-colocated generation.

Run async GRPO with the Single Controller so rollout and training overlap, reducing GPU idle time on a Megatron backend.

```
$ uv run examples/run_grpo_single_controller.py --config
examples/configs/grpo_math_1B_megatron_single_controller.yaml
```

Enable streaming and choose the `ready_first` sampler so the trainer starts as soon as 8 prompt groups finish and reuses late-arriving groups instead of dropping them.

```yaml
async_rl:
  min_groups_for_streaming_train: 8
  sampler:
    name: ready_first
    max_staleness_versions: 1
```

— Names every config key, sampler option, and a runnable command v0.8.0

02 **Weight refit transports for policy-to-generation sync**

IMPROVED 90

NeMo RL adds multiple weight-refit transports between policy and generation workers: colocated `IPC`, `NCCL`, sparse delta, and `NIXL` transports, including `NIXL` checkpoint-engine refit to update non-colocated vLLM generation workers from policy workers. Setting `policy.generation.refit_transport: nccl_reshard` sends FFN and expert weight slices shard-to-shard over NVLink without assembling the full tensor on any rank, cutting non-colocated refit time on large MoE models by up to 5× (reported as DeepSeek-V3 refit dropping from 15s to 3s). Setting `policy.generation.refit_transport: nixl` with `refit_cfg.nixl.backend_name: UCX` sends full weights over a UCX/RDMA fabric, 1.54× faster than broadcast NCCL refit on DeepSeek-V3.

— Exact config keys and measured speedups for each transport v0.8.0

03 **New rollout engines: TensorRT-LLM and Dynamo** NEW 78

NeMo RL adds TensorRT-LLM and Dynamo as new rollout engines alongside vLLM, SGLang, and Megatron Inference. Set `policy.generation.backend: trtllm` with `trtllm_cfg.async_engine: true` to use TensorRT-LLM for colocated or non-colocated rollout with Megatron and Automodel training, exposed via an OpenAI-compatible HTTP server for NeMo Gym agentic rollouts such as SWE; managed Dynamo generation is also supported as a rollout engine.

Switch to TensorRT-LLM as the rollout engine for a non-colocated run to serve NeMo Gym agentic workloads via its OpenAI-compatible HTTP server.

```yaml
policy:
  generation:
    backend: trtllm
    trtllm_cfg:
      tensor_parallel_size: 1
      async_engine: true
      gpu_memory_utilization: 0.6
```

— Config key and flag given, but Dynamo support described only briefly v0.8.0

04 **NVFP4/W4A4 quantized training on Blackwell** NEW 68

NeMo RL adds `W4A4` policy training and per-token vLLM rollout on NVIDIA Blackwell GPUs, with end-to-end `NVFP4` training plus quantized GRPO and distillation via NVIDIA ModelOpt.

— Names hardware, precision formats and ModelOpt but no config shown product docs

THINNER COVERAGE BELOW

05 **New training recipes: audio-visual intent GRPO and Nemotron 3 Ultra**

NEW 55

NeMo RL adds the Audio-Visual Intent GRPO workflow to train Qwen2.5-Omni-7B with GRPO on audio-visual intent recognition datasets and evaluate on Daily-Omni, and adds post-training of Nemotron 3 Ultra with RLVR, teacher training, and MOPD stages on GB200 NVL72 hardware.

— Names models and stages but no commands or configs product docs

06 **OpenTelemetry observability via nemo-lens** NEW 50

NeMo RL adds OpenTelemetry traces and `rl.*` metrics via `nemo-lens`, including span groups, configuration options, vLLM tracing, and an OTLP export stack.

— Names the tool and metrics but no setup steps given product docs

07 **LoRA fine-tuning across Automodel and Megatron Core** NEW 40

NeMo RL adds parameter-efficient fine-tuning with LoRA across both the Automodel and Megatron Core backends.

— Bare statement with no flags or config detail product docs

WAS THIS USEFUL? 👍 👎

◆ Speech & Voice AI

# Parakeet ASR Server

Feedback ⓘ v0.10.0 NOTES ↗ RANK

Parakeet ASR Server transcribes audio on CPUs or GPUs using NVIDIA Parakeet ONNX models through a Whisper-compatible API.

> Parakeet ASR Server adds configurable worker concurrency and multi-GPU/MIG slice distribution for transcription workloads, alongside two breaking changes to GPU flag handling and CUDA startup behavior.

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

## 01 Multi-GPU and MIG slice distribution via -gpu-devices

95

BREAKING

`-gpu-devices` takes a comma-separated list of GPU indices or MIG slices (e.g. `0,1,2`) to spread transcription workers across multiple GPUs, or the equivalent `PARAKEET_GPU_DEVICES` environment variable for container use; each worker gets its own encoder and decoder so no resources are shared between requests. This replaces the previous single-GPU `-gpu-device` flag, so existing invocations using `-gpu-device` must be updated.

Run the server across three GPUs with two workers per slice to maximise throughput on a multi-GPU host.

```
$ ./parakeet -gpu cuda -gpu-devices 0,1,2 -workers 6
```

Use environment variables to configure GPU device selection in a container environment where CLI flags are inconvenient.

```
$ PARAKEET_GPU=cuda PARAKEET_GPU_DEVICES=0,1 ./parakeet -workers
4
```

Distribute four workers across two GPUs (or two MIG slices) so each transcription has its own encoder and decoder and GPU races are eliminated.

```
$ ./parakeet -gpu cuda -gpu-devices 0,1 -workers 4
```

— Flag, env var, mechanism and migration note all named with runnable commands. v0.10.0

02 **Configurable worker concurrency with VRAM sizing** NEW

80

The `-workers` flag (default `4`) bounds the entire transcription pipeline — audio in to text out — to a configurable concurrency limit, so a busy server queues cleanly instead of over-allocating. Capacity can be planned using documented per-worker VRAM footprints: approximately 670 MB per worker with int8 quantization, or 2.5 GB with fp32.

Run four simultaneous transcriptions on a CPU-only server so load queues gracefully instead of causing out-of-memory crashes.

```
$ ./parakeet -workers 4
```

— Names flag, default, and concrete VRAM figures with a runnable example. v0.10.0

03 **CUDA startup now fails fast instead of silent CPU fallback**

60

BREAKING

GPU transcription is enabled with `-gpu cuda` or `PARAKEET_GPU=cuda`; the server now refuses to start rather than silently falling back to CPU when CUDA is unavailable or the driver is incompatible, so deployments relying on the prior silent fallback will fail on startup.

— Names exact flags and new startup behavior but no runnable example given. product docs

└──▷ **BREAKING ON UPGRADE**

! When CUDA is selected but unavailable (missing library or wrong driver), the server now refuses to start instead of falling back to CPU — deployments that relied on the silent CPU fallback will fail on startup.

! The `-gpu-device` flag (single GPU) is replaced by `-gpu-devices` (comma-separated list); existing invocations using `-gpu-device` must be updated.

WAS THIS USEFUL? 👍 👎

◆ Image & Video Generation

## Runway

Feedback ⓘ App Studio NOTES ↗ RANK

AI video and image generation platform with models for text-to-video, image-to-video and editing, available in a web app and through an API.

> Runway introduced App Studio, a plain-language builder that generates, brands and hosts custom creative apps powered by Runway's models.

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

01 **App Studio plain-language app builder** NEW 35

App Studio lets users describe a creative app in plain language, then generates, brands, and hosts it using Runway's models and the user's own files.

— Describes what it does but no concrete UI path or config App Studio

WAS THIS USEFUL? 👍 👎

**AGENT**

◆ AI Agent Frameworks

## Arcade

Feedback ⓘ updated · Nov 7, 2026 NOTES ↗

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

Arcade adds per-execution API key and gateway tracking, database migration controls, OpenFGA least-privilege provisioning, and Excel conflict-handling.

Each tool execution now records which API key and gateway it ran through, reported alongside the execution's other details.

Control database migrations at upgrade time with the `engine.database.migrations` config key (set to `false` to skip automatic migration on `helm upgrade`).

Control OpenFGA database provisioning with `rbac.openfga.database.create: false` to skip running `createdb` against the postgres maintenance database, allowing a DBA-created database with a least-privilege role.

UPDATED INTEGRATION Handle file-creation conflicts in Excel's upload and create calls with the new `conflict_behavior` parameter, accepting `fail`, `rename`, or `replace` in place of the old overwrite flag.

NEW INTEGRATION GitHub QueryIssues and QueryPullRequests

NEW INTEGRATION GitHub ListRepositoryConversationComments

NEW INTEGRATION Slack DeleteFiles

NEW INTEGRATION Linear ListRecentComments

NEW INTEGRATION Microsoft Teams ListMeetingArtifacts

UPDATED INTEGRATION Linear ListIssues creator filter and new fields

UPDATED INTEGRATION Google Drive SearchFiles pagination

RETIRED INTEGRATION Figma ListProjects and ListFiles removed

└──▷ BREAKING ON UPGRADE

! The Figma toolkit removes `ListProjects` and `ListFiles`; callers using either tool must migrate before upgrading.

! Excel's upload and create calls replace the plain overwrite flag with `conflict_behavior` (`fail`, `rename`, `replace`); existing calls that passed the overwrite flag will break and must be updated.

! Microsoft Office toolkits move to broader file scopes; action is required on upgrade (see upgrade notes).

! `rbac.openfga.database.create: false` no longer runs `createdb` against the postgres maintenance database; operators relying on the previous automatic

database creation must pre-create the database before upgrading.

! Word now refuses to silently replace an existing same-named document on create; the default behavior changes to refusing the create and reporting the conflict — callers that expected silent replacement must pass `conflict_behavior`.

---

WAS THIS USEFUL? 👍 👎

# CopilotKit OpenBot

Feedback ⓘ v0.3.2 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 v0.3.2 release adds a self-updating Windows/Mac Desktop with database backup-and-rollback, an in-window browser teaching flow, on-device narration, and a personal-edition installer, alongside several breaking changes to Helm configuration, desktop environment variables, and tool access for Claude Agent SDK and AG-UI/Mastra Bots.

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

## 01 Helm chart learning-model and migration timeout keys

BREAKING

90

The Helm chart now controls Automatic Learning's model via `config.learning.model` instead of `intelligence.learning.agent.model`; an upgrade that still passes `intelligence.learning.agent.model` is refused. Its `intelligence.learning.queue.migration.activeDeadlineSeconds` now defaults to 900 (15 minutes) instead of 300, to allow more time for PostgreSQL readiness; an upgrade run with `--reuse-values` keeps the old 300-second limit unless `intelligence.learning.queue.migration.activeDeadlineSeconds=900` is passed explicitly.

Pin the Automatic Learning model in a Helm upgrade without getting refused for using the old key.

```
$ helm upgrade openbot ./chart --set config.learning.model=gpt-4o
```

— Exact Helm keys, defaults, and a runnable upgrade command v0.3.2

## 02 Self-updating Desktop with backup and rollback

NEW

80

OpenBot Desktop on Mac and Windows now checks for updates at launch and every six hours, showing an 'OpenBot 0.3.2 is available.' bar with an Update and restart button that fetches updates from `downloads.copilotkit.ai` (falling back to GitHub). On Windows, Update and restart downloads and verifies the installer and applies it with no prompts while keeping the existing installation and settings; the upgrade process backs up the database beforehand and automatically rolls back files and data on any failure, showing per-step status messages ('Stopping OpenBot.', 'Backing up OpenBot's database.', 'Downloading OpenBot Y.', 'Installing OpenBot Y.', 'Starting OpenBot Y.',

'Putting OpenBot's database back.'). Windows installations from v0.3.1 or earlier have no built-in updater and must be upgraded by hand once before automatic updates work.

— Full mechanism and messages given but no CLI/config surface v0.3.2

03 **brand.yaml tenant.former_ids for renamed deployments** NEW

75

The `brand.yaml` configuration file now accepts `tenant.former_ids` so a deployment previously set up under a different package name keeps its Bots, conversations, and deployment name. A personal edition set up with `start.sh` under v0.3.1 must add `tenant.former_ids` with the old package name since the server package is now called `openbot`.

Retain Bots, conversations, and deployment name after the package rename from a custom name to `openbot` in a self-hosted personal edition.

yaml

```yaml
tenant:
  former_ids:
    - personal
```

— Named config key with migration example to run v0.3.2

04 **Desktop environment variable rewrite on Start** BREAKING

70

OpenBot Desktop now rewrites `~/OpenBot/.env` on Start, adding `OPENBOT_PUBLIC_URL`, `OPENBOT_THIS_MACHINE_CONTAINER_HOST`, `COMPUTER_LOOPBACK_RELAYS`, `INTELLIGENCE_API_URL`, and `INTELLIGENCE_GATEWAY_WS_URL`, and removing `OPENBOT_BROWSER_URL`, `OPENBOT_BROWSER_CONTAINER_URL`, `BROWSER_PORT`, and `OPENBOT_BROWSER_ABSENT`. The server no longer reads `OPENBOT_BROWSER_ABSENT`, so any automation that sets it must remove it.

— Every env var named but no explanation of why they changed v0.3.2

05 **Optional supervisor-less server config** NEW

70

Setting `COMPUTER_SUPERVISOR_CONTAINER_URL` empty in `docker-compose.yml` gives the server no supervisor; leaving it unset defaults to `http://supervisor:4300`.

Disable the supervisor in a Compose deployment so the server starts with no supervisor sidecar.

env

```env
COMPUTER_SUPERVISOR_CONTAINER_URL=
```

— Exact env var and file with default value given v0.3.2

THINNER COVERAGE BELOW

06 **Claude Agent SDK Bots restricted to OpenBot tools** BREAKING 55

Claude Agent SDK Bots no longer receive Claude Code's own tools (Bash, Read, WebSearch, Agent, etc.); they now act only through the tools OpenBot provides.

— Names the removed tools but no migration path given v0.3.2

07 **On-device speech narration in Desktop** NEW 55

Narration in Desktop now transcribes speech on-device: via the Speech framework on Mac (with a one-time permission prompt) and Windows' own speech recognition on Windows, so voice never leaves the computer.

— Names both platform mechanisms but no setting to toggle v0.3.2

08 **AG-UI/Mastra Bots lose copilotkit.ui field** BREAKING 50

Bots registered by their own AG-UI or Mastra address no longer receive `forwardedProps.copilotkit.ui`; agents that read that field will no longer find it.

— Names the exact field but no alternative or fix described v0.3.2

09 **Personal edition installer for Desktop** NEW 45

OpenBot Desktop now installs the personal edition, sharing one Bot computer and starting with the Personal Assistant, and downloads three fewer images than the team edition.

— Describes behaviour but no config to adjust it v0.3.2

10 **Starter Bot suggested tasks** NEW 40

The starter Bot now offers five suggested tasks in the first new conversation: connect an app, build an interactive answer, search the web, read a page in its browser, and add a task to a demo checklist.

— Lists all five tasks but no mechanism behind them v0.3.2

11 **In-window browser for Teach a Bot** IMPROVED 35

Teach a Bot now runs the browser inside the OpenBot Desktop window rather than opening a separate Chrome or Edge window.

— Clear before/after but no further mechanism detail v0.3.2

12 **New Bot creation with no managed Bot** NEW 35

New Bot now works on any deployment with no managed Bot, including Desktop, allowing a Bot to be created from instructions alone.

— States capability but no steps or config shown v0.3.2

└──▷ **BREAKING ON UPGRADE**

! The Helm chart no longer accepts `intelligence.learning.agent.model`; set `config.learning.model` instead — an upgrade that still passes `intelligence.learning.agent.model` is refused.

! The Helm chart's `intelligence.learning.queue.migration.activeDeadlineSeconds` default rises from 300 to 900; an upgrade run with `--reuse-values` keeps the old 300-second limit — pass `intelligence.learning.queue.migration.activeDeadlineSeconds=900` explicitly or drop the flag.

! The server no longer reads `OPENBOT_BROWSER_ABSENT`; remove it from any automation that sets it.

! Claude Agent SDK Bots no longer receive Claude Code's own tools (Bash, Read, WebSearch, Agent, etc.); they now act only through the tools OpenBot provides.

! Bots registered by their own AG-UI or Mastra address no longer receive `forwardedProps.copilotkit.ui`; agents that read that field will no longer find it.

! A personal edition previously set up with `start.sh` under v0.3.1 must add `tenant.former_ids` to `brand.yaml` with the old package name to retain its Bots, conversations, and deployment name, because the server package is now called `openbot`.

! Windows installations of OpenBot Desktop from v0.3.1 or earlier have no built-in updater and must be upgraded by hand once before automatic updates work.

---

WAS THIS USEFUL? 👍 👎

## CrewAI

Feedback ⓘ 1.15.37 NOTES ↗ RANK

Open-source Python framework for building multi-agent workflows using role-based AI agent Crews and event-driven Flows.

> CrewAI added spreadsheet reading for agents and support for scaffolding standalone agent projects outside a full Crew setup.

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

01 **Spreadsheet reading for agents** NEW 45

Agents can now read spreadsheet files via `crewai-files`, giving crews direct access to structured tabular data.

— Names the mechanism but lacks formats or limits 1.15.37

02 **Standalone agent project scaffolding** NEW 35

CrewAI now supports scaffolding and loading standalone agent projects, enabling agents to be developed and run outside a full Crew setup.

— No command or flag named for scaffolding 1.15.37

WAS THIS USEFUL? 👍 👎

# Microsoft Semantic Kernel

Feedback ⓘ dotnet-1.82.0

NOTES ↗ RANK

Semantic Kernel provides an open-source SDK for building AI agents and integrating language models into C#, Python, and Java applications.

> Semantic Kernel's dotnet-1.82.0 release focuses on cleanup and hardening: it drops Python 3.10 support, removes the legacy .NET Milvus and Chroma connectors in favor of their new locations, and consolidates file-access and HTTP-address handling in .NET as breaking changes, alongside smaller OpenAPI parser and termination-strategy improvements.

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

## 01 Python version support updated to drop 3.10, add 3.14

55

BREAKING

Semantic Kernel drops Python 3.10 support and extends CI coverage through Python 3.14.

— Clear version numbers but no migration steps given. dotnet-1.82.0

## 02 Python AggregatorTerminationStrategy composability IMPROVED

55

The Python `AggregatorTerminationStrategy` now derives from `TerminationStrategy`, making it composable with other termination strategies.

— Names exact class and base class, no usage example. dotnet-1.82.0

## 03 Milvus and Chroma .NET connectors removed BREAKING

45

The .NET connector for Milvus is removed after migration to the new connector location, and the .NET connector for Chroma is removed after migration to the new connector location; existing code referencing the old connector packages must migrate.

— Named connectors removed but no new location path given. dotnet-1.82.0

## 04 File access and HTTP address boundary hardening in .NET

40

BREAKING

Semantic Kernel hardens file-access and HTTP-address boundaries in .NET; the `[BREAKING] Consolidate file access` change alters how file access is structured in .NET, so existing code using the old file-access API will break.

— Breaking change named but exact mechanism unspecified. dotnet-1.82.0

## 05 OpenAPI parser path-item and non-method key support NEW

30

The OpenAPI parser now adds support for path-item parameters and non-method keys.

— Thin description with no usage example or schema detail. dotnet-1.82.0

### └──▷ BREAKING ON UPGRADE

! The .NET Milvus connector package has been removed; migrate to the new connector location.

! The .NET Chroma connector package has been removed; migrate to the new connector location.

! The .NET `[BREAKING] Consolidate file access` change alters how file access is structured in .NET; existing code using the old file-access API will break.

WAS THIS USEFUL?  

# Google ADK

Feedback ⓘ v2.14.0 NOTES ↗ RANK

The Agent Development Kit is Google's code-first Python toolkit for building, evaluating, and deploying agents, with built-in tools, sessions, and deployment to Vertex AI.

> Google ADK v2.14.0 adds multi-run evaluation with configurable parallelism, semantic tool-argument matching, Microsoft Teams and OpenAI Decisions integrations, and vector search/TTL in MongoDbMemoryService, alongside chunked large-file streaming across execution environments and six breaking changes touching session services, skill scripts, enum tool arguments, and live-run WebSocket handling.

└──▷ **WHAT SHIPPED · 15 FEATURES** most completely described first ? what's the number?

## 01 Stricter kwargs validation in VertexAiSessionService

BREAKING

90

`VertexAiSessionService.create_session` now rejects extra keyword arguments other than `display_name`, `ttl`, `expire_time`, `labels`, and their camelCase aliases, raising `ValueError` (HTTP 400 from `adk api_server`) for keys such as `http_options`, `session_state`, `sessionId`, and `wait_for_completion`; use `state=` and `session_id=` instead.

— Exact accepted/rejected keys and replacement params named. v2.14.0

## 02 RunSkillScriptTool drops command argument

BREAKING

90

`RunSkillScriptTool` no longer accepts a `command` argument when used with a `SkillToolset` environment — the model must now pass `args`, `short_options`, and `positional_args`; only `.py`, `.sh`, and `.bash` scripts are supported (others return `UNSUPPORTED_SCRIPT_TYPE`), and calls that pass `command` now return `INVALID_ARGUMENTS`.

— Exact replacement arguments and error codes documented. v2.14.0

## 03 Enum parameters now passed as enum members

BREAKING

88

`FunctionTool` parameters annotated `Enum`, `Optional[Enum]`, or `list[Enum]` now receive enum members instead of raw strings. Code comparing with `arg == "value"`, using string-keyed dict lookups, `match` on strings, `json.dumps`, or string methods on those arguments breaks; compare against members, use `.value`, or switch to `StrEnum`.

— Full before/after behavior and migration guidance given. v2.14.0

04 **Multi-run evaluation with configurable parallelism** NEW

85

Evaluation cases can now run multiple times concurrently via `MultiRunConfig`, the `adk eval --num_runs` flag, and `--parallelism` flag, then combine scores or pass rates across runs into a single stable metric.

Run an eval set 5 times concurrently to get stable pass-rate averages across non-deterministic agent responses.

```
$ adk eval \
    contributing/samples/evaluation/home_automation_agent \

contributing/samples/evaluation/basic_criteria/home_automation.ev
alset.json \
    --config_file_path
contributing/samples/evaluation/basic_criteria/eval_config.json \
    --num_runs 5 \
    --parallelism 5
```

— Concrete command, flags, and score-combination mechanism given. v2.14.0

05 **Microsoft Teams integration via TeamsRunner** NEW 80

A new `TeamsRunner` class connects an ADK agent to Microsoft Teams chats and channels; install the integration with the new `teams` extra (`pip install "google-adk[teams]"`).

Connect an ADK agent to Microsoft Teams so it responds in chats and channels.

```python
pip install "google-adk[teams]"

from google.adk.integrations.teams import TeamsRunner
from my_agent import root_agent

runner = TeamsRunner(agent=root_agent)
runner.run()
```

— Runnable install command and code example provided. v2.14.0

06 **Vector search and TTL in MongoDbMemoryService** IMPROVED

73

`MongoDbMemoryService` adds vector search retrieval with `retrieval_mode="vector"` or `retrieval_mode="auto"`, and memory expiration via `memory_ttl_days`.

— Config parameters named but no usage example given. v2.14.0

---

07 **Chunked large-file streaming across environments** IMPROVED

73

`BaseEnvironment.read_file_stream` streams large files in chunks across `LocalEnvironment`, `E2BEnvironment`, and `DaytonaEnvironment`, letting `ReadFileTool` read line ranges without loading whole files into memory. `AgentEngineEnvironment` now also reads and writes large files in chunks and reports commands running over 570 seconds as timed out before the service deadline.

— Named classes and method but no runnable example. v2.14.0

---

08 **WebSocket close codes for /run_live** BREAKING

70

`/run_live` now closes the socket with code 1007 on a malformed frame and code 1000 when the run ends — reuse of an idle socket across runs no longer works; reconnect for each new live run.

— Exact close codes and required client behavior given. v2.14.0

---

09 **External storage path for eval results** NEW

65

The `adk eval` command gains an `--eval_storage_uri` flag to store evaluation output files outside a read-only agents directory, such as a GCS bucket path.

Save eval output to a writable GCS path when the agents directory is read-only (e.g. a container image).

```
$ adk eval \
    contributing/samples/evaluation/home_automation_agent \

contributing/samples/evaluation/basic_criteria/home_automation.ev
alset.json \
    --eval_storage_uri gs://my-bucket/eval-results/
```

— Flag and runnable example given, but little mechanism detail. v2.14.0

---

10 **Semantic tool argument matching in evals** NEW

65

Evaluations can compare tool call arguments using the new `ToolArgsMatcher`, supporting exact, case-insensitive, numeric-tolerance, or LLM-backed semantic checks.

— Matching modes named but no usage example provided. v2.14.0

11 **PostgreSQL JSON column type change in DatabaseSessionService** 65

BREAKING

`DatabaseSessionService` creates new PostgreSQL JSON columns as `json` instead of `jsonb`; queries outside ADK that rely on jsonb-only operators or GIN indexes will not work on new tables.

— Names the column type change but gives no migration steps. v2.14.0

12 **FirestoreSessionService event payload format change** BREAKING 60

`FirestoreSessionService` stores event payloads in a new format; sessions written by v2.14.0 cannot be read by v2.13.0 or earlier after a rollback.

— States incompatibility but no migration path given. v2.14.0

THINNER COVERAGE BELOW

13 **OpenAI Decisions API classifier** NEW 55

The new `OpenAIDecisionsClassifier` class classifies workflow and tool-gating questions against the OpenAI Decisions API.

— Class named but no usage example or config shown. v2.14.0

14 **Hot-reload flag for adk run** NEW 50

A new `--reload_agents` flag on `adk run` enables hot-reloading agents during development.

— Flag named but no further workflow description. v2.14.0

15 **A2A agent card generation CLI** NEW 45

A new `adk generate_agent_card` command generates an A2A agent card from the CLI.

— Only command name given, no usage detail or example. v2.14.0

└──▷ **BREAKING ON UPGRADE**

! `FunctionTool` parameters annotated `Enum`, `Optional[Enum]`, or `list[Enum]` now receive enum members instead of raw strings — `arg == "value"`, string-keyed dict lookups, `match` on strings, `json.dumps`, and string methods on those arguments no longer work; compare against members, use `.value`, or switch to `StrEnum`.

! `VertexAiSessionService.create_session` now rejects extra keyword arguments other than `display_name`, `ttl`, `expire_time`, `labels`, and their camelCase aliases, raising `ValueError` (HTTP 400 from `adk api_server`) for keys such as `http_options`, `session_state`, `sessionId`, and `wait_for_completion`; use `state=` and `session_id=` instead.

! `RunSkillScriptTool` no longer accepts a `command` argument when used with a `SkillToolset` environment — the model must now pass `args`, `short_options`, and `positional_args`; only `.py`, `.sh`, and `.bash` scripts are supported (others return `UNSUPPORTED_SCRIPT_TYPE`); calls that pass `command` now return `INVALID_ARGUMENTS`.

! `FirestoreSessionService` stores event payloads in a new format; sessions written by v2.14.0 cannot be read by v2.13.0 or earlier after a rollback.

! `DatabaseSessionService` creates new PostgreSQL JSON columns as `json` instead of `jsonb`; queries outside ADK that rely on jsonb-only operators or GIN indexes will not work on new tables.

! `/run_live` now closes the socket with code 1007 on a malformed frame and 1000 when the run ends — reuse of an idle socket across runs no longer works; reconnect for each new live run.

---

WAS THIS USEFUL? 👍 👎

## Microsoft Agent Framework

Feedback ⓘ dotnet-1.28.0

NOTES ↗

The Microsoft Agent Framework builds, orchestrates, and deploys agents and multi-agent workflows in Python and .NET, consolidating AutoGen and Semantic Kernel.

Microsoft Agent Framework adds postpone-task support in A2A hosting, CodeAct host-tool diagnostic redaction, MCP session scoping to invocations, and Anthropic citation streaming.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch dotnet-1.28.0 https://github.com/microsoft/agent-framework.git
# already have the repo? check out this version:
$ git checkout dotnet-1.28.0
```

Adds postpone-task support in A2A hosting for .NET.

Now redacts host-tool diagnostics by default in CodeAct for both .NET and Python.

Now scopes declarative MCP sessions and workflow MCP clients to individual invocations.

Now binds AG-UI workflow agent executors to one authorization scope in Python.

Now hardens MCP catalog discovery behavior in Python.

+4 more

NEW SUPPORTED PLATFORM Redis integration on Python 3.15

RETIRED SUPPORTED PLATFORM Python 3.10 support removed

└──▷ BREAKING ON UPGRADE

! Python: Declarative MCP sessions are now scoped to invocations — sessions that were previously shared across calls will no longer persist between them.

! .NET: Workflow MCP clients are now scoped to invocations — any shared client state across workflow invocations will be lost.

! Python: AG-UI workflow agent executors are now bound to a single authorization scope.

! Python: MCP catalog discovery behavior is hardened — previously permissive discovery patterns may now be rejected.

! Python: Foundry Local integration now requires SDK 2.3.1.

- ! CodeAct now redacts host-tool diagnostics by default in both .NET and Python.
- ! .NET: A2A agent creation and payload handling has been simplified — existing construction code may need updating.
- ! .NET: Allowed HTTP hosts are now required — setups without explicit allowed-hosts configuration will fail on upgrade.
- ! Workflow span names are now aligned with OpenTelemetry conventions in both Python and .NET.
- ! Python 3.10 is no longer supported — workloads must upgrade to Python 3.11 or later.

WAS THIS USEFUL? 👍 👎

## CopilotKit

Feedback ⓘ v1.85.0 NOTES ↗ RANK

CopilotKit connects AI agents to application interfaces with shared state, tools, and generative UI components.

> CopilotKit extends Generative UI support to Angular and adds analytics context forwarding for Intelligence runs.

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

01 **Native Generative UI for Angular** NEW 40

CopilotKit now supports native Generative UI in Angular, extending a capability previously available only in React.

— States scope and prior limitation but no implementation detail v1.85.0

02 **Analytics context forwarding for Intelligence runs** NEW 30

CopilotKit now forwards analytics context for Intelligence runs, making run data available to analytics pipelines.

— Describes benefit but no named fields, API, or config v1.85.0

WAS THIS USEFUL? 👍 👎

# NVIDIA Object-Oriented Agents (NOOA)

Feedback ⓘ

v0.0.11 NOTES ↗ RANK

Object-Oriented Agents provides a Pythonic framework for building AI agents.

> Object-Oriented Agents v0.0.11 adds a Decision API for routing and classification with lightweight decision models, explicit per-client media input controls, a guided `nooa connect` setup wizard, and shared request-rate pacing, alongside several breaking changes around media configuration and Windows support.

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

## 01 Media input restrictions for ResponsesClient

BREAKING

85

The `media_inputs` parameter on `ResponsesClient` in `nooa.unifiedllm` declares exactly which media types a model receives, accepting `'image'`, `'pdf'`, `'audio'`, and `'video'`. This is a breaking change: existing model configurations must declare `media_inputs` to keep sending media, since omitting it now stops media from being sent.

Restrict a model client to only receive image and PDF inputs, so attempts to send audio or video are rejected before the request is sent.

python

```python
from nooa.unifiedllm import ResponsesClient

model = ResponsesClient("openai/your-model", media_inputs=
["image", "pdf"])
```

— Names param, accepted values, module, and migration impact v0.0.11

## 02 Decision API with DecideStrategy

NEW

80

`DecideStrategy` can be applied to any agentic method via @strategy(DecideStrategy()), with `decision_model=<model>` declared on the Agent subclass, enabling routing, classification, or scoring with a specialized decision model instead of invoking a full generative LLM.

Use a fast decision model to route support messages to billing or technical without invoking a full generative LLM.

```python
from nooa import Agent, strategy
from nooa.strategies import DecideStrategy
from typing import Literal

class SupportRouter(Agent, decision_model=jev):
    @strategy(DecideStrategy())
    async def route(self, message: str) -> Literal["billing",
"technical"]:
        """Choose the team to handle this message."""
        ...
```

— Names the exact decorator and class param with a worked example `v0.0.11`

03 **Guided `nooa connect` setup wizard** NEW 75

Run `uv run nooa connect` to walk through provider selection, API key entry, model settings, media checks, and saving, replacing manual model configuration.

Walk through the interactive model setup wizard to configure a provider, set media support, run live checks, and save the alias for use in agent code.

```
$ uv run nooa connect
```

— Runnable command with full wizard steps listed `v0.0.11`

04 **Native Windows support gated by platform check** BREAKING 75

Native Windows is refused by default on `import nooa` and the `nooa` command; set `NOOA_SKIP_PLATFORM_CHECK=1` to allow the tested subset, or use WSL 2 for features requiring POSIX.

— Names exact env var and the behavior change `v0.0.11`

05 **File size cap for ShellTools** NEW 65

`ShellTools` whole-file reads and replacements are capped by the `max_file_bytes` setting, which defaults to 8 MiB.

— Names config key and default, no usage example `v0.0.11`

THINNER COVERAGE BELOW

06 **Application-scoped request-rate pacing** NEW 55

Object-Oriented Agents adds application-scoped request-rate pacing with shared cooldowns derived from provider Retry-After headers, working alongside existing concurrency limits.

— Explains mechanism but no config surface named v0.0.11

---

07 **LLM call tracing helper for custom clients** NEW 40

The new journal_llm_call() helper records LLM calls made from custom model clients.

— Names the function but no usage detail given v0.0.11

---

08 **Explicit context_window required with API key** BREAKING 35

Passing an explicit API key to a registry alias now requires an explicit `context_window` to be set.

— States the requirement but no further mechanism given v0.0.11

└──▷ **BREAKING ON UPGRADE**

! Existing model configurations must declare `media_inputs` to continue sending media; omitting it causes media to stop being sent.

! Native Windows is refused by default on `import nooa` and the `nooa` command; set `NOOA_SKIP_PLATFORM_CHECK=1` to allow the tested subset, or use WSL 2.

! Passing an explicit API key to a registry alias now requires an explicit `context_window` to be set.

---

WAS THIS USEFUL?  

# lexmount Moli

Feedback ⓘ v1.1.22 NOTES ↗ RANK

Best headless browser for AI agents. Lite, Fast, High-Compatibility. Built in Rust

> Moli added child document response interception with authentication challenge support via `fetch`, and native WebAuthn credential signal validation.

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

01 **Child document interception with auth challenges** NEW 40

Moli can now intercept child document responses and authentication challenges via `fetch` using a new interception capability.

— Names the mechanism but no config or API detail v1.1.22

02 **Native WebAuthn credential signal validation** NEW 35

Adds `webauthn` credential signal validation through native bindings.

— Names the feature flag but lacks usage detail v1.1.22

WAS THIS USEFUL? 👍 👎

# cotal

Feedback ⓘ v0.147.0 NOTES ↗ RANK

The open standard for agent coordination

> cotal's latest releases focus on identity and chat ergonomics — stable, @mention-reachable chat names with strict validation and breaking changes for older sessions — alongside automatic dashboard credential minting for `cotal up` / `cotal service start web` and tighter restart-safety requirements.

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

## 01 Stable @mention-reachable chat names

BREAKING — 95

`cotal chat` now opens a session under a stable name reachable via @mention, set with `--name` or persisted via `cotal config set name <name>` (cleared with `cotal config unset name`); without either, it falls back to the IdP handle on a user-auth space or the OS login on an open/static space. Names must be 1–128 characters using letters, digits, `_`, `.`, or `-`, and must start and end with a letter, digit, or `_`; `--name` no longer accepts spaces, `@`, or other disallowed characters. Sessions from older cotal versions, or whose IdP provides no valid chat name, have no handle on record and `cotal chat` refuses with `AUTH_REQUIRED` until re-login or a name is saved, and older cotal versions reading preferences after `name` is set fail with `defaults has an unknown key "name"` until upgraded or the key is removed.

Persist a stable @mention-reachable name for your chat sessions so agents can address you consistently across restarts.

```
$ cotal config set name alice_ops
```

Open a chat session under an explicit name when you want to override the stored preference for a single session.

```
$ cotal chat --name alice_ops
```

Remove a saved name preference so older cotal versions stop failing on the unknown key.

```
$ cotal config unset name
```

— Names exact flags, config commands, validation rules and breaking error states v0.147.0

02 **Automatic dashboard credential minting at cotal up** NEW

85

`cotal up` now mints a read-only dashboard credential stored in `web.creds`, and the `web` service preset automatically uses it as the `--creds` default, so `cotal service start web` can be run with no flags.

— Names the exact config key, flag default, and zero-flag command `v0.146.0`

03 **Login handle recorded for userStatus** IMPROVED

65

`cotal login` now records a handle, chosen as the IdP's preferred username, else email local-part, else display name, and `userStatus` returns it as `login.name`.

— Names the login mechanism and the returned field but gives no runnable example `v0.147.0`

04 **Restart safety now requires proof of work closure** BREAKING

60

Automatic restarts are refused unless proof of local work closure is provided; the built-in PTY runtime cannot supply this proof, so explicit `supervise` requests are rejected before launch, and crash recovery (including recovery of older supervised records) is also refused without authoritative work closure.

— Explains the new restriction and affected subsystems but offers no workaround command `v0.146.0`

THINNER COVERAGE BELOW

05 **TUI channel messaging and threaded replies** NEW

30

The terminal UI (TUI) now supports writing to channels and agents directly, and replying to a specific picked message within a TUI conversation thread.

— Describes capability only in prose with no commands or UI path `v0.147.0`

### BREAKING ON UPGRADE

! `--name` no longer accepts names that contain a space, an `@`, or any character outside letters, digits, `_`, `.`, and `-`, or names that do not start and end with a letter, digit, or `_`.

! A user-auth session created by an older version of cotal, or whose IdP provides no valid chat name, has no handle on record; `cotal chat` will refuse with `AUTH_REQUIRED` until you sign in again (`cotal login`) or save a name with `cotal config set name <name>`.

! An older cotal reading preferences after the `name` key is saved will fail every preferences read with `defaults has an unknown key "name"` until it is upgraded or the name is removed with `cotal config unset name`.

! Automatic restarts are refused unless proof of local work closure is provided; the built-in PTY runtime cannot supply this proof, so explicit `supervise` requests are rejected before launch. Pi crash recovery and recovery of older supervised records are also refused without authoritative work closure.

WAS THIS USEFUL? 👍 👎

## agentlas-os

Feedback ⓘ os-v1.2.87 NOTES ↗ RANK

Agent OS: keep specialist agents in a hub, spin up a temporary orchestrator per task. Local-first, works with any model.

> Agentlas OS tightened its update and prompting internals: it now automatically prunes stale engine versions after an update, and filters code map excerpts so they only attach to prompts that reference real indexed code.

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

### 01 Automatic cleanup of old engine versions after update

IMPROVED 70

After an update, the engine removes old engine versions from the per-user runtime directory, keeping the current version, any version referenced by a host config or running process, anything placed in the last 15 minutes, one previous release for rollback, and (after a manual rollback) the highest newer complete release. It removes at most five versions per run and writes one line to stderr when it cannot determine what is in use.

— Detailed retention logic and limits, but runs automatically with no user command.

os-v1.2.87

THINNER COVERAGE BELOW

### 02 Smarter code map excerpt filtering for agent prompts IMPROVED

58

The per-turn code map excerpt is now attached to a prompt only when the prompt names a real indexed code identifier or file, excluding vendored, dependency, and mirrored plugin paths, and replacing control characters in file names. Subagent prompts follow the same rule.

— Explains filtering rules and scope but offers no user-facing action. os-v1.2.87

WAS THIS USEFUL? 👍 👎

## yoagent

Feedback ⓘ v0.27.2 NOTES ↗

yoagent is a Rust agent loop that streams from 7 LLM protocols and runs tools until done, from a laptop to a Cloudflare Worker.

yoagent adds `PriceEntry::with_limits` to store context window and max output on price-table entries, and ModelConfig::bedrock_anthropic_id(id) to resolve Bedrock model IDs to Anthropic IDs for custom pricing.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.27.2 https://github.com/yologdev/yoagent.git
# already have the repo? check out this version:
$ git checkout v0.27.2
```

Look up the Anthropic model ID behind a Bedrock model ID or inference profile ID using ModelConfig::bedrock_anthropic_id(id), so hosts can apply Anthropic pricing rates to Bedrock usage.

Price-table entries now carry `context_window` and `max_output` fields via `PriceEntry::with_limits`, and the bundled snapshot applies these limits even when pricing is disabled.

NEW SUPPORTED PLATFORM Claude Opus 4.7, 4.6 and 4.5 limits and prices

WAS THIS USEFUL? 👍 👎

# openclacky

Feedback ⓘ v1.5.23 NOTES ↗ RANK

openclacky is a token-efficient open-source AI agent framework.

> OpenClacky v1.5.23 adds durable delegated work via a new secretary extension, support for running multiple instances of the same IM channel platform, several Web UI file/media/editor upgrades, and persistent composer drafts across reloads.

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

01 **Web UI file, media and editor handling** IMPROVED 70

The Web UI now shows linked local media inline, opens Markdown images in the lightbox, adds file actions to the image lightbox toolbar, adds an open-with submenu to file context menus, and syncs scrolling between the Markdown editor and preview.

— Names every UI surface changed, giving clear navigation points v1.5.23

THINNER COVERAGE BELOW

02 **Composer draft persistence across reloads** NEW 40

Composer text drafts are now persisted per session, so in-progress messages survive page refreshes.

— Explains behaviour and scope but no config surface v1.5.23

03 **Scheduled task usability improvements** IMPROVED 35

Scheduled tasks now offer starter examples when none exist yet, and show a readable title instead of the ASCII name.

— Describes two thin UI tweaks to the same area v1.5.23

04 **Icons and detail pages for builtin extensions** IMPROVED 35

Builtin extensions now have icons and full detail pages.

— Names the UI addition but not its location or scope v1.5.23

05 **Secretary extension for durable delegated work** NEW 30

Adds durable delegated work and a new secretary extension for handling background tasks.

— Names the extension but no mechanism or config shown v1.5.23

06 **No-vision inference for custom models** IMPROVED 30

OpenClacky now infers no-vision capability for custom models based on known model names.

— Explains mechanism briefly but no examples of model names v1.5.23

07 **Multi-instance IM channel support** NEW 25

OpenClacky now supports running multiple instances of the same IM channel platform in parallel.

— States capability without naming platforms or setup steps v1.5.23

08 **Session naming in desktop notifications** IMPROVED 20

Desktop notifications now name the session they relate to.

— Single-line description with no further mechanism v1.5.23

09 **Refreshed image and video model lineup** IMPROVED 20

Refreshes the OpenClacky image and video model lineup.

— No specific models or counts named v1.5.23

WAS THIS USEFUL?  

◆ MCP Servers & Tooling

# gflow-cli

Feedback ⓘ v0.94.0 NOTES ↗ RANK

Drive Google Flow from the command line: Veo video and Imagen images, scripted, batched and pipeline-ready. Ships an MCP server so coding agents can drive it too, giving you and your agents full programmatic control over generation.

> gflow-cli's v0.94.0 release introduces local usage tracking, letting users inspect generation counts and telemetry state entirely on-device.

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

01 **Local usage stats and telemetry commands** NEW 75

gflow-cli now tracks local usage counters, stored on-device with no remote transmission in this release. Users can check usage statistics with the new `gflow data stats` command, and inspect telemetry collection state with the new `gflow telemetry status` and `gflow telemetry show` commands.

Review how many generations you have run locally without leaving the terminal.

```
$ gflow data stats
```

Confirm that telemetry is local-only and see what has been recorded before enabling remote reporting.

```
$ gflow telemetry status
gflow telemetry show
```

— Names exact commands and storage mechanism, no limits or retention details given

v0.94.0

WAS THIS USEFUL? 👍 👎

# opentakeoff

Feedback ⓘ mcp-v0.9.105 NOTES ↗ RANK

0) PDF takeoff for construction & flooring — the first engine an AI agent drives natively over MCP, not bolted on. Rooms measured on the wall faces, symbol sweeps, materials + quantities with full provenance, built for preconstruction.

> opentakeoff's mcp-v0.9.105 release adds per-line estimate pricing via the MCP `set_pricing` command, extends MCP access to trade suites with a new canvas Trade view, and introduces an in-canvas AI agent chat alongside a split-screen reference pane and refreshed estimate view and premium dock UI.

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

01 **Per-line estimate pricing via `set_pricing`** NEW 75

The MCP `set_pricing` command prices every line of an estimate in a single call; a bordered run's derived line is now optional.

Price every line of an estimate in one MCP call without manually entering unit costs.

```
$ set_pricing
```

— Names exact MCP command and gives a runnable example. mcp-v0.9.105

02 **In-canvas AI agent chat** NEW 60

A new Agent chat window lets you talk to your own AI model from inside the canvas; it streams replies and adds a Conversation memory setting.

— Names setting and behaviour, but no API/command given. mcp-v0.9.105

THINNER COVERAGE BELOW

03 **Expanded estimate view line items** IMPROVED 45

Estimate view now shows a line for every quantity a condition carries, including assemblies, with an always-on watermark.

— Describes behaviour change but no config or command. mcp-v0.9.105

04 **Trade suites accessible over MCP** NEW 40

Trade suites can now be accessed over MCP and viewed in a new canvas Trade view.

— Names surface but no command or config detail. mcp-v0.9.105

05 **Premium dock tool tray** IMPROVED 40

The Premium dock adds a tool tray (⋯) and cleans up the drawing bar.

— Minor UI change, concrete icon named but thin detail. mcp-v0.9.105

06 **Split-screen reference pane** NEW 35

A read-only reference-pane split-screen lets you compare plan sheets side-by-side while working.

— UI feature described briefly, no exact path given. mcp-v0.9.105

WAS THIS USEFUL? 👍 👎

## novamira

Feedback ⓘ v1.12.15 NOTES ↗

MCP server that gives AI agents full access to WordPress through PHP execution and filesystem operations

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v1.12.15 https://github.com/use-
novamira/novamira.git
# already have the repo? check out this version:
$ git checkout v1.12.15
```

NEW INTEGRATION Gemini Desktop connector

WAS THIS USEFUL? 👍 👎

## gram

Feedback ⓘ server@2.31.0 NOTES ↗ RANK

Gram provides MCP servers and AI-agent tools from APIs with access controls and observability.

> Gram replaced CLI authentication with a PKCE one-time code flow, added an MCP Gateway column to the staff coverage matrix, removed external MCP server support from deployments and the admin dashboard, and rolled out GitHub Actions Pull Request job support, a downloadable Claude observability plugin, and a string of agent dashboard, telemetry and Code Mode refinements.

└──▷ **WHAT SHIPPED · 16 FEATURES** most completely described first ? what's the number?

### 01 PKCE-based authentication for speakeasy auth IMPROVED

87

`speakeasy auth` now signs in via a PKCE one-time code flow instead of an API key callback: the CLI sends a code challenge with the browser login, the dashboard returns a short-lived code bound to that challenge, and the CLI redeems it with a verifier that never leaves the process, preventing callback-swap and eavesdropping attacks. Use `key_scope: producer` on `cliAuth.authorize` to request the flow, which requires org:admin permission (matching the permission for creating a producer key in the dashboard) and records the key in the audit log; single-use state tokens are carried on the callback URL so stray or forged callbacks are rejected.

— Full mechanism and named flag/permission, but UI entry point not shown cli@0.19.4 dashboard@0.145.0

### 02 Project overview and chat list telemetry consolidation

85

BREAKING

`telemetry.getProjectOverview` gains an `include_usage` flag returning per-user and per-agent usage breakdowns in one response, while dropping the `comparison`, `metrics_mode`, `resolved_chats`, `failed_chats`, `active_users_count`, `top_users`, and `llm_client_breakdown` fields (and no longer running the queries behind them). `chat.list` gains a `pinned_first` flag so recent chats load in one call instead of two.

Load the project overview with usage breakdowns included in a single call, avoiding chained telemetry and user-search requests.

```
$ telemetry.getProjectOverview({ include_usage: true })
```

Load a project's pinned and recent chats in one call rather than two separate requests.

```
$ chat.list({ pinned_first: true })
```

— Named flags and dropped fields with runnable example calls `server@2.31.0`

03 **Security headers on chat attachment responses** IMPROVED

85

Chat attachment responses from `assets.serveChatAttachment` and `assets.serveChatAttachmentSigned` now include `Content-Security-Policy: default-src 'none'; style-src 'unsafe-inline'; sandbox` and `X-Content-Type-Options: nosniff` headers, with Content-Disposition controlling inline display for raster images, audio, plain text, CSV, Markdown, and JSON, and forcing download for HTML, SVG, XML, and PDF.

— Exact header values and endpoint names with full format list `dashboard@0.145.0`

04 **Pull Request job for GitHub Actions agents** NEW

78

GitHub Actions agent setup offers a Pull requests job choice that writes the exact rule `repo:<owner>/<repo>:pull_request`, generates a workflow running only on `pull_request` events (opened, synchronize, reopened, ready for review), and skips fork pull requests. Available where the `workload-platform-github-actions` feature flag is on.

— Exact rule and event triggers named, gated behind a feature flag `server@2.31.0`

05 **Org-wide MCP server activity endpoint** NEW

78

A new `telemetry.getOrgMcpServerActivity` endpoint returns recent MCP server activity for every readable project in one request, replacing per-project individual requests on agent identity pages.

— Named endpoint with clear replacement behaviour, no example call `server@2.31.0`

06 **External MCP server support removed** BREAKING

73

The `external_mcps` field on deployment creation and the `upsert_external_mcps` field on evolve are removed; deployments no longer accept or carry external MCP servers, and evolving or redeploying a deployment produces one without them. The External MCP migration page and its admin API are also removed from the admin dashboard.

— Named fields and removed admin surface, migration path unclear `admin@0.11.0` `server@2.31.0`

07 **Agent dashboard activity, sessions and tool logs updates** NEW

73

The Activity tab now records `agent:connection_attach` and `agent:connection_detach` events alongside the existing `remote_session:attach` and `remote_session:detach` entries. Members with `agent:write` on an agent can now list and open that agent's sessions read-only via its Sessions panel. Tool Logs now include local tool calls by default, with a Show local tools switch on the agent overview applying to both the Tool calls card and Tool logs together.

— Named events and permission scope, UI path given but not exact `dashboard@0.145.0`

---

08 **MCP Gateway column in coverage matrix** NEW 70

The staff admin coverage matrix gains an MCP Gateway column measuring tool calls, identity (person-bound vs. managed-agent-bound), and policy enforcement via `risk_findings` rows with a non-empty `mediation_surface`. Cells can carry a new `na` status for capabilities the gateway architecturally cannot report (it brokers tool calls, not model calls, and shadow servers are reached around it), and each cell now has a `unit` field naming what it counts — the gateway column in tool calls, agent surfaces in sessions.

— Named fields and mechanism, but internal staff-only tool limits reader action

`admin@0.11.0` `server@2.31.0`

---

09 **Reconnect message in Code Mode tools.connect** IMPROVED 70

In Code Mode, `tools.connect` now returns a reconnect message when a gateway server gains a sign-in requirement after being approved as needing none, instead of a sign-in link.

— Named API and precise before/after behaviour change `dashboard@0.145.0`

---

10 **Client name reporting to MCP authorization servers** NEW 65

An opt-in setting on remote session clients sends the connecting MCP client's self-reported name to the provider as `speakeasy_client_name` on `/authorize`, off by default and only sent for anonymous sessions. A companion setting on provider clients sends that name to the authorization server so it can label each connection.

— Named field and defaults given, but toggle location not specified `dashboard@0.145.0` `server@2.31.0`

---

11 **Claude observability plugin download** NEW 60

Download an organization's Claude observability plugin to upload to claude.ai, enabling cloud sessions to be recorded. Each download is keyed to the downloading user and replaces their previous one; the shared org install key is refused.

— Clear behaviour and constraints, but no exact UI location given `server@2.31.0`

---

12 **Per-organization onboarding customization** NEW 60

Onboarding customization is now accessible from each organization's sidebar and Overview, with wizard step configuration, shared template application, and a separate recorded stack section in a tenant-scoped workspace.

– Names UI entry points but no config keys or API given `admin@0.11.0`

THINNER COVERAGE BELOW

13 **GitHub Actions workflow pinned to device agent v0.1** IMPROVED 55

GitHub Actions setup workflow now uses the device agent action's `v0.1` tag instead of a pre-release commit, so jobs pick up new 0.1.x releases without workflow edits; speakeasyd `version` is left unset so the action installs the latest stable release.

– Names the tag and setting but low reader-facing mechanism `dashboard@0.145.0`

14 **Gateway instructions name reachable servers** IMPROVED 55

Gateway instructions now open with a line naming the servers the caller can reach (within 250 characters, with a count when servers do not fit), so agents that load tools on demand know the gateway's scope before searching.

– Concrete character limit given but no config surface to act on `dashboard@0.145.0`

15 **Authenticated Remote MCP connection test** NEW 50

A new setup connection test initializes a Remote MCP connection and discovers tools using the signed-in user's saved authorization, without executing any tools.

– Describes behaviour but no named endpoint or UI path `server@2.31.0`

16 **Unavailable tools shown in Inspect tab** NEW 45

The Inspect tab now lists every tool an MCP server offers and shows which ones a user's access does not include under Not available.

– Clear UI behaviour but thin on mechanism `dashboard@0.145.0`

└──▷ **BREAKING ON UPGRADE**

! The `external_mcps` field on deployment creation and the `upsert_external_mcps` field on evolve are removed; evolving or redeploying a deployment now produces one without any external MCP servers its source had.

! `telemetry.getProjectOverview` no longer returns `comparison`, `metrics_mode`, or the summary fields `resolved_chats`, `failed_chats`, `active_users_count`, `top_users`, and `llm_client_breakdown`; callers reading these fields will receive nothing.

! The External MCP migration page and its admin API are removed from the admin dashboard.

WAS THIS USEFUL?  

## context7

Feedback ⓘ ctx7@0.10.0 NOTES ↗ RANK

Context7 maintains up-to-date documentation and context for AI agents through an MCP server.

> Context7 0.10.0 adds first-class setup support for the Pi coding agent and T3 Code's providers, and switches its VS Code/Cursor config editing to an in-place method that preserves existing file formatting.

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

### 01 Pi coding agent setup and removal NEW 90

`ctx7 setup --pi` writes the MCP server config to `~/.pi/agent/mcp.json` (or `.pi/mcp.json` with `--project`), adds the Context7 rule to an existing `AGENTS.override.md` or `CLAUDE.md`, and installs the skill in `~/.agents/skills`. It respects the `PI_CODING_AGENT_DIR` environment variable, and prints the `pi mcp login context7` step when `--oauth` is passed; `ctx7 remove` supports the same `--pi` flag to undo the setup.

Register Context7 with the Pi agent globally, using OAuth so Pi can authenticate against the MCP server.

```
$ ctx7 setup --pi --oauth
```

Register Context7 with the Pi agent scoped to the current project directory instead of the global agent config.

```
$ ctx7 setup --pi --project
```

— Names exact flags, config paths and env var ctx7@0.10.0

### 02 T3 Code provider support NEW 66

The new `--t3` flag on `ctx7 setup` and `ctx7 remove` configures T3 Code's Claude Code and Codex providers for Context7.

Configure T3 Code's Claude Code and Codex providers for Context7 in one step.

```
$ ctx7 setup --t3
```

— Clear flag and scope, little mechanism detail ctx7@0.10.0

THINNER COVERAGE BELOW

### 03 In-place editing of VS Code and Cursor MCP configs IMPROVED

59

`ctx7 setup` and `ctx7 remove` now edit `.vscode/mcp.json` and `.cursor/mcp.json` in-place, preserving comments and formatting instead of rewriting the file through `JSON.stringify`.

— Names exact files and before/after behavior `ctx7@0.10.0`

WAS THIS USEFUL? 👍 👎

## davinci-resolve-mcp Feedback ⓘ v4.19.2 NOTES ↗ RANK

davinci-resolve-mcp is an MCP server integration for DaVinci Resolve Studio.

> davinci-resolve-mcp adds verification fields to its audio embedding capability reports so callers can confirm whether CLAP audio embedding actually succeeded rather than just assuming availability.

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

### 01 Audio embedding verification fields and probe function NEW

80

Embedding capability reports now include `audio.verified` (set true only after a real embedding call completes in the current process) and `audio.last_error` (the most recent failure message, cleared on next success). The new probe_audio_backend() function executes a real audio and text embedding call to populate these fields, loading (and downloading if needed) the model on first use.

Update to the latest version and restart your MCP client to get the new audio verification fields.

```
$ npx davinci-resolve-mcp@latest sync
```

— Names exact fields and function, explains mechanism and trigger `v4.19.2`

WAS THIS USEFUL? 👍 👎

# agentcad

Feedback ⓘ v0.7.2 NOTES ↗ RANK

CAD CLI and MCP server for AI agents

> agentcad v0.7.2 adds recovery and workflow-control primitives for long-running CAD operations: idempotency keys for safely retrying runs and imports, staged model promotion, and resumable background inspection, alongside retention controls for saved inspection responses.

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

01 **Idempotency keys for run and import recovery** NEW 85

Reuse `--idempotency-key` on `run` and `import` to recover a lost response without re-executing the underlying script or file; a saved run result can be read with `run --idempotency-key KEY --status` even after the source script has changed or been removed.

Re-fetch the result of a run that timed out or whose response was lost, without executing the script again.

```
$ agentcad run script.py --label first --idempotency-key KEY --status
```

— Exact flags and a runnable command given v0.7.2

02 **Background inspection with resumable checkpoints** NEW 80

Return from an import before inspection finishes using `--inspection-background` on `import`; the response includes actions to check progress, cancel, or resume from saved checkpoints without importing the CAD file again.

Start a CAD import and let inspection run in the background so the agent can continue work while the file is processed.

```
$ agentcad import model.step --inspection-background
```

— Named flag with a runnable example command v0.7.2

03 **Staged model promotion via agentcad promote** NEW 70

Save a candidate model without replacing the accepted version using `--stage` on a run, then accept it later with `agentcad promote` without re-running its source.

— Named flag and command but no example shown v0.7.2

04 **Inspection response budgets and retention cleanup** NEW

60

Set text and image budgets for inspection replies and retrieve the full saved response when needed; saved full responses default to 30 days and 64 MiB per build directory, with cleanup preserving CAD files, geometry IDs, retry records, and resumable jobs.

— Concrete limits given but no flag or key named v0.7.2

THINNER COVERAGE BELOW

05 **Cloud setup guide for hosted MCP URLs** IMPROVED

15

The cloud setup guide now includes direct connections to hosted MCP URLs.

— Bare documentation mention, no detail v0.7.2

WAS THIS USEFUL?

◆ Agent Memory

# claude-mem

Feedback ⓘ v15.11.0 NOTES ↗ RANK

claude-mem gives coding agents persistent memory by capturing session activity, compressing it with a model, and injecting relevant context into later sessions.

> claude-mem's only release this window adds per-screen installer telemetry that tracks setup progress across nine named screens while respecting user privacy preferences.

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

01 **Installer telemetry with privacy-safe schema and setup-attempt tracking**

NEW 69

The telemetry schema accepts only known screen IDs, actions, error codes, timestamps, and random attempt/visit IDs; prompts, filenames, URLs, credentials, and free-form error text are excluded, and the schema respects `DO_NOT_TRACK`. Setup branches are tracked accurately — Home and existing Pro configurations skip checkout, history import returns to Finish setup, and reduced-motion setup skips the launch animation — and a fresh setup-attempt ID is passed through the installer pairing when telemetry is enabled, connecting setup progress to the authorized account flow in the admin report.

— Names env var and branch logic but no admin report UI shown v15.11.0

WAS THIS USEFUL? 👍 👎

◆ Browser & Computer Use

# Skyvern

Feedback ⓘ v1.0.57 NOTES ↗ RANK

Skyvern automates browser-based workflows with vision models and agents, completing form-driven tasks on sites it has not seen before.

> Skyvern's v1.0.57 release adds email forwarding and inbox recovery for 2FA codes, Word .docx SOP ingestion for Copilot, scoped API key permission enforcement, and AWS Secrets Manager storage for self-hosted credentials, alongside smaller RBAC, workflow-secret, and UI improvements.

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

01 **AWS Secrets Manager storage for Credentials page** NEW

60

On self-hosted AWS deployments, secrets entered on the Credentials page are now stored in AWS Secrets Manager.

— Names the exact storage backend and deployment target v1.0.57

02 **Email forwarding and inbox for 2FA codes** NEW 60

Skyvern adds email forwarding setup and an inbox UI to receive and recover forwarded email 2FA codes, and rejects reuse of revoked forwarding addresses.

— Describes the flow and a specific safeguard, no UI path given v1.0.57

THINNER COVERAGE BELOW

03 **Scoped API key permission enforcement** NEW 45

API keys are now enforced to their declared scope, so a key can only perform the actions within the permissions it was issued for.

— Names the mechanism but no scope list or example v1.0.57

04 **Word .docx SOP ingestion in Copilot** NEW 45

Copilot can now take Word `.docx` SOPs from the Upload SOP dialog and chat and read them.

— Names the dialog and file format but no further detail v1.0.57

05 **Agent settings reorganized into folding groups** IMPROVED

40

Agent settings are now grouped into six folding groups with change summaries.

— Gives a concrete count but no navigation path v1.0.57

06 **Copilot test run cost and failure diagnostics** IMPROVED

35

Copilot now shows what each test run cost and explains why a test browser did not start.

— States what changed but no numbers or UI location v1.0.57

07 **RBAC roles seeded from Clerk memberships** NEW

35

RBAC roles are now seeded from live Clerk memberships.

— Names Clerk integration but no further mechanism v1.0.57

08 **Demonstrate single block after prior blocks run** NEW

35

Users can now demonstrate one block of an existing agent after its earlier blocks have already run.

— Describes the capability but no workflow example v1.0.57

09 **Code blocks read downloaded document text** NEW

30

Code blocks can now read the text of a downloaded document.

— Bare description with no mechanism or scope v1.0.57

10 **Write-only inline workflow secrets** BREAKING

30

Inline workflow secrets are now write-only.

— One-line change with no explanation of impact v1.0.57

WAS THIS USEFUL?  

◆ Applied AI Agents

# nanobot

Feedback ⓘ v0.4.1 NOTES ↗ RANK

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

> nanobot's v0.4.1 release adds a file-context graph with previews and snapshots, unifies syntax highlighting across code views, and tightens up OpenAI API configuration while flagging legacy importers for removal.

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

01 **OpenAI apiType conversion removed** BREAKING 85

The temporary OpenAI `apiType` / `api_type` conversion introduced in 0.4.0 is no longer accepted after this release; users must save the migrated `api` configuration and replace old environment variables before upgrading past 0.4.1.

— Clear migration step naming config keys and env vars to replace. v0.4.1

02 **Legacy JSONL and proxy importers deprecated** DEPRECATED 65

Legacy JSONL and proxy importers remain available throughout 0.4.x but are scheduled for removal in 0.5.0; users should retain original files and migration backups now.

— States removal version and concrete action to preserve backups. v0.4.1

THINNER COVERAGE BELOW

03 **Unified syntax highlighting across code views** IMPROVED 55

nanobot now shares syntax highlighting across code views, source previews, and read-only diffs, with complete source copying and plain-text fallbacks.

— Names affected surfaces but no flag or command to invoke. v0.4.1

04 **File-context graph with previews and snapshots** NEW 35

nanobot adds a file-context graph with file previews, local Markdown links, and per-conversation context snapshots.

— Named from summary only, no mechanism or UI path given. v0.4.1

└──▷ **BREAKING ON UPGRADE**

! The temporary OpenAI `apiType` / `api_type` conversion introduced in 0.4.0 is no longer accepted after this release; save the migrated `api` configuration and replace old environment variables before upgrading past 0.4.1.

! Legacy JSONL and proxy importers remain available throughout 0.4.x but are scheduled for removal in 0.5.0; retain original files and migration backups now.

---

WAS THIS USEFUL? 👍 👎

## DEPLOY

◆ Inference Providers & GPU Clouds

# Crusoe Cloud

Feedback ⓘ updated · Nov 7, 2026 NOTES ↗ RANK

Crusoe Cloud provides GPU infrastructure and managed inference for AI training and deployment.

> Crusoe Cloud shipped Kubernetes control plane audit logging for CMK v2 clusters and extended Telemetry Conduit to forward metrics and logs to external OTLP-over-HTTP endpoints.

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

01 **Kubernetes control plane audit logs for CMK v2 clusters** NEW

80

Audit logs from the Kubernetes API server and other control plane components can now be queried via the `LogsQL` API, viewed in `Command Center > Managed Logs`, or streamed out via Telemetry Conduit — available for every CMK v2 cluster and enabled by default with no setup required.

— Names API and UI path but not exact query syntax snapshot-20261107

02 **OpenTelemetry forwarding via Telemetry Conduit** IMPROVED

70

Telemetry Conduit now forwards metrics and logs, including CMK control plane audit logs, to any OTLP-over-HTTP endpoint such as a third-party observability platform or SIEM, with per-configuration destination, credential, and filter settings.

— Describes mechanism but no exact config keys named snapshot-20261107

WAS THIS USEFUL? 👍 👎

## Anyscale

Feedback ⓘ Version 0.26.115 release notes NOTES ↗ RANK

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

> Anyscale added an `ended_at` timestamp to job and service status and listing commands, letting users track termination times for jobs and services.

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

01 **`ended_at` field for jobs and services** NEW 81

Anyscale adds a new `ended_at` field, readable via `anyscale job status`, `anyscale service status`, anyscale.job.status(), and anyscale.service.status(), which remains unset until the job or service terminates. It's also available in bulk listings via `anyscale job list --v2` and `anyscale service list -o json`, as well as through anyscale.job.list() and anyscale.service.list() in the SDK.

Audit termination times across all services in bulk for SLA reporting.

```
$ anyscale service list -o json | jq '[.[] | {name: .name,
ended_at: .ended_at}]'
```

— Names exact commands/SDK calls but lacks broader mechanism detail

Version 0.26.115 release notes

WAS THIS USEFUL? 👍 👎

## AWS Amazon Bedrock

Feedback ⓘ

Added Kimi K3.1 Pro - Highspeed (2026-11-04) BLOG ↗

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

NEW INTEGRATION Kimi K3.1 Pro - Highspeed (Moonshot AI)

WAS THIS USEFUL? 👍 👎

◆ AI / LLM Gateways

# OpenRouter

Feedback ⓘ updated · Nov 7, 2026 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 rebuilt its Auto Router and introduced a new Auto Code router, both using an eleven-factor live evaluation of pricing, provider health, benchmarks, and cache state to select the best model for a request.

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

01 **Rebuilt Auto Router with eleven-factor selection** IMPROVED

53

OpenRouter's Auto Router now weighs eleven factors per request — including live pricing, provider health, benchmarks, and cache state — to select which model handles a request.

— Names selection factors but no API, flag, or config surface. snapshot-20261107

02 **Auto Code router for coding workloads** NEW

53

OpenRouter adds Auto Code, a new router specialized for coding work that applies the same eleven-factor live evaluation (live pricing, provider health, benchmarks, and cache state) used by the Auto Router, but tuned to route coding requests.

— Explains mechanism and purpose but no concrete API or model ID to invoke it.

snapshot-20261107

WAS THIS USEFUL? 👍 👎

## dario

Feedback ⓘ v6.22.3 NOTES ↗ RANK

One local endpoint that routes Cursor, Cline, Aider, Claude Code and the Agent SDK through your Claude and ChatGPT subscriptions, with failover and pooling and zero runtime dependencies.

> dario shipped a byte-offset HTTP endpoint for remote log access and shared ChatGPT seat pooling across replicas, the latter introducing new lock-service storage requirements and alias conventions.

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

### 01 Shared ChatGPT seat pooling across replicas BREAKING

93

`--pool-shared-accounts` shares ChatGPT seats across all replicas behind a load balancer, syncing seats through the lock service using the same rules already applied to Claude accounts: newer login wins, then later expiry, and removals propagate to every replica. `dario codex remove` publishes a seat removal to all replicas before the local removal to keep the registry consistent. This is a breaking change: ChatGPT refresh tokens are now stored on the lock service when pooling is active and must be guarded like the accounts directory, and replicas sharing the same alias (default `altman-1` on every replica) will keep only the newer login on upgrade, so distinct aliases must be assigned to each replica beforehand.

Start dario with seat pooling enabled so ChatGPT seats are visible and usable across all replicas behind a load balancer.

```
$ dario --pool-shared-accounts
```

Remove a ChatGPT seat and propagate the removal to every replica immediately.

```
$ dario codex remove
```

— Named flag, subcommand, sync rules, and explicit upgrade/security caveats. v6.22.0

### 02 HTTP endpoint for request log access NEW

87

The new `GET /analytics/requests` endpoint serves the per-request log file over HTTP, returning content by byte offset (whole lines only) and a pagination mark to detect file rotation; it answers 404 when logging is disabled. This lets a sidecar or remote container read the log without a shared filesystem.

Read the request log from a sidecar or remote container without sharing the credentials folder — paginate forward by passing the last byte offset as a query parameter.

```
$ curl -s 'https://<host>/analytics/requests?offset=0' \
  -H 'Authorization: Bearer <token>'
```

— Named endpoint, method, and runnable curl example with pagination mechanism. v6.22.3

### └──▷ BREAKING ON UPGRADE

! ChatGPT refresh tokens are now stored on the lock service when `--pool-shared-accounts` is active — guard the lock service with the same care as the accounts directory.

! Replicas holding different ChatGPT subscriptions under the same alias keep only the newer login on upgrade; assign distinct aliases to each replica before upgrading (the default alias is `altman-1` on every replica).

---

WAS THIS USEFUL? 👍 👎

# Mozilla AI Otari

Feedback ⓘ v0.40.0 NOTES ↗ RANK

Open-source, OpenAI-compatible LLM gateway you run yourself. One endpoint for 40+ providers, with virtual keys, budgets, and usage tracking.

> Otari v0.40.0 adds access-rule controls for stored MCP servers, a SearXNG adapter for the any-search subsystem, and self-service budget visibility, while introducing several breaking changes around organization-key naming, audio pricing, and MCP server configuration.

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

01 **Access rules for stored MCP servers** BREAKING 65

MCP stored servers now support access rules for API keys and users, with a `read-tools` preset to control which tools each key or user may call. This is now required: existing MCP server configurations without access rules will need to be updated on upgrade.

— Names the read-tools preset but no API/CLI surface given v0.40.0

02 **SearXNG adapter in any-search subsystem** NEW 60

Adds the SearXNG adapter to the `any-search` subsystem, routing all SearXNG tool calls through it.

— Names the adapter and subsystem but no config example v0.40.0

03 **Deprecation of web_search_url and web_search_engines settings** DEPRECATED 60

The settings `web_search_url` and `web_search_engines` are now deprecated in favor of the `any-search` adapters; a startup warning names what their removal changes.

— Names both deprecated keys and the replacement system v0.40.0

THINNER COVERAGE BELOW

04 **Pass-through endpoint and search cost tracking** IMPROVED 50

API pass-through endpoints now report what each call bills, tracking cost per pass-through route. Provider-run searches are now priced at the model's own rate.

— Describes two billing changes without naming specific endpoints v0.40.0

05 **Unpriced speech requests refused, audio cost on headers** 50

BREAKING

Unpriced speech requests are now refused instead of silently passed through; audio cost is instead returned on response headers.

— Explains before/after behaviour but no header name given v0.40.0

06 **Request-prefix identification for custom-endpoint organization keys**

45

BREAKING

Organization keys named with a custom endpoint are now identified by a request prefix — existing naming conventions for such keys will break on upgrade.

— States the change but not the exact prefix format v0.40.0

07 **Self-service budget visibility and dashboard limits** NEW

40

An API key can now read its own budget and usage directly, without admin credentials. Budget token and request limits can now be set from the dashboard, without editing config files.

— Describes capability but no endpoint or UI path named v0.40.0

08 **Per-repository guardrail opt-out** NEW

35

Lets a person turn their own guardrails off in one repository from the hook layer.

— Thin description, no mechanism or config key named v0.40.0

└──▷ BREAKING ON UPGRADE

! Organization keys named with their own endpoint are now identified by a request prefix — existing naming conventions for such keys will break on upgrade.

! Unpriced speech requests are now refused; audio cost is returned on response headers instead of being silently passed through.

! Stored MCP servers now require access rules for keys and users — existing MCP server configurations without access rules will need to be updated.

WAS THIS USEFUL? 👍 👎

## OmniRoute

Feedback ⓘ v3.8.53 NOTES ↗

OmniRoute puts hundreds of model providers behind one endpoint, with quota-aware fallback and prompt compression.

OmniRoute v3.8.53 adds per-API-key provider pinning, quota session recovery, Entra ID auth, agentic-stall and anti-avalanche failover guards, Gemini prompt-cache tracking, team billing cost centers, and OTel cost attribution.

└──▷ GET THIS VERSION ▾

Docker

```
$ docker pull ghcr.io/cheaper-inference/omniroute:3.8.53
```

git

```
$ git clone --branch v3.8.53 https://github.com/cheaper-
inference/OmniRoute.git
# already have the repo? check out this version:
$ git checkout v3.8.53
```

published under the project's GitHub account · checked 7 Nov Wrong command?

Add per-API-key ordered `preferredConnections` so each project can pin a primary provider account with quota/cooldown-aware automatic fallback to the next eligible account.

Enable opt-in quota session recovery by sending `x-omniroute-quota-recovery: auto` on a non-streaming chat turn to park a quota-blocked request (encrypted, 24 h) and poll its result at `GET /api/v1/quota-recoveries/{id}` instead of failing hard.

NEW INTEGRATION Authenticate Claude Code and any client to `/v1/*` with a Microsoft Entra ID token instead of a static API key, mapping Entra security groups onto existing OmniRoute key groups; includes a dependency-free sign-in helper at `scripts/cli/omniroute-sso.mjs`.

Enable opt-in agentic-stall failover with `OMNIROUTE_AGENTIC_STALL_FAILOVER=1` (default OFF) so a combo member that returns a stop with no tool call fails over to the next target.

Enable opt-in anti-avalanche failover guard with `OMNIROUTE_AVALANCHE_PROTECTION_ENABLED=true` (off by default).

+12 more

NEW INTEGRATION ainetcafe OpenAI-compatible provider

NEW INTEGRATION Kimi (Moonshot) web search provider

NEW INTEGRATION WhyCodes CLI Agent

NEW INTEGRATION GitHub Copilot OAuth connection

NEW INTEGRATION Freebuff browser login OAuth connection

NEW INTEGRATION Antigravity Manager account export import

NEW INTEGRATION sing-box embedded TPROXY sidecar service

UPDATED INTEGRATION Grok Build OAuth — video, speech, transcription and image generation

NEW INTEGRATION Bifrost sidecar fast-path

WAS THIS USEFUL? 👍 👎

◆ ML Platforms & SDKs

## Hugging Face

Feedback ⓘ A New Main Search UI NOTES ↗ RANK

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

> Hugging Face replaced its main search bar with a command palette for jumping to models, datasets, and Spaces.

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

01 **Command palette search UI** NEW 70

The main search bar is replaced by a command palette, opened by clicking the search bar or pressing `⌘K` or `Ctrl+K`, that lets users navigate to any model, dataset, or Space with arrow-key navigation and quick actions to create new resources.

Jump to a model, dataset, or Space without leaving the keyboard.

> 📍 Press Ctrl+K (or ⌘K on Mac) to open the command palette, type a resource name, navigate with arrow keys, and press Enter to go directly to it.

— Names exact shortcut and UI behaviour, runnable steps given. A New Main Search UI

WAS THIS USEFUL? 👍 👎

## emisar

Feedback ⓘ v0.57.0 NOTES ↗

Emisar lets AI agents securely run infrastructure actions through MCP, with approval controls that protect production systems.

emisar adds least-privilege host-access recipes, SSO form fixes, `emisar doctor` replacing `emisar state check-dispatch-log`, and `emisar-mcp connect` multi-server support.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.57.0 https://github.com/AndrewDryga/emisar.git
# already have the repo? check out this version:
$ git checkout v0.57.0
```

Reach the identity provider from sign-in, MFA, invitation, and identity-link forms through the handoff page, so `form-action 'self'` no longer blocks SSO flows.

Inspect dispatch state with `emisar doctor`, which now replaces `emisar state check-dispatch-log`.

Keep other MCP servers in Codex and Grok configs when running `emisar-mcp connect`.

Least-privilege python-app and terraform-readonly recipes now grant a group instead of root, refuse a folder whose group is root, and report what code runs as the runner.

Pending-approval emails now go to both the person who started the run and those who can approve it.

NEW VULNERABILITY CHECK Image scan evidenced exceptions for zlib and ntpsec

NEW INTEGRATION Medusa pack for Cassandra

RETIRED INTEGRATION git-local, cloud-init, php-fpm older pack versions retired

└──▷ BREAKING ON UPGRADE

! `emisar state check-dispatch-log` is removed; use `emisar doctor` instead.

! The `DATABASE_SSL_CACERTFILE` environment variable is retired.

! The release-cookie cutover toggle and the legacy checkout preflight are retired.

! The Windows MCP installer drops the `-ConnectAll` flag.

! git-local versions before 0.1.12, cloud-init versions before 0.1.20, and php-fpm versions before 0.1.17 are retired and must be upgraded.

WAS THIS USEFUL?  

## skillhub

Feedback ⓘ v0.2.26 NOTES ↗ RANK

Self-hosted, open-source agent skill registry for enterprises. Publish & version skill packages, govern with RBAC and audit logs, deploy on-premise with Docker or Kubernetes.

> SkillHub v0.2.26 adds opt-in SMTP email notifications, a default-off OTLP/HTTP business-event logging pipeline, configuration-only frontend analytics via Umami and DataFinder, richer skill presentation with versioned summaries, and a new review center — alongside a required database migration.

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

01 **Required database migration on upgrade** BREAKING 75

Upgrading applies Flyway migrations `V120` – `V121` ; back up PostgreSQL and skill storage before upgrading. Schema rollback requires restoring the matching pre-upgrade backup.

– Exact migration versions and rollback procedure given v0.2.26

02 **OTLP/HTTP business-event Logs pipeline** NEW 70

Adds a default-off `OTLP/HTTP` business-event Logs pipeline covering publication, prepared downloads, review and promotion decisions, and supported lifecycle changes. Export is bounded and best-effort, not an audit or billing guarantee.

– Names protocol and covered events, no exact config key v0.2.26

03 **Frontend analytics via Umami and DataFinder** NEW 70

Supports configuration-only frontend analytics with Umami and DataFinder adapters, same-origin proxy routes, and collection controls. Collection is disabled by default and independent of application startup.

– Names adapters and proxy routes but no exact config key v0.2.26

04 **SMTP email notifications** NEW 65

Adds optional SMTP email notification delivery with per-category immediate or digest preferences, separate email verification, persistent delivery status, and controlled retries. Both platform delivery and user preferences are disabled by default.

– Describes mechanism well but no config key or flag named v0.2.26

05 **Review center for skill approvals** NEW 60

Adds a review center with unified review filtering, task details, comparison against the published baseline, and required comparison failures that block approval.

— Clear UI starting point, no exact navigation path given v0.2.26

THINNER COVERAGE BELOW

06 **Versioned skill summaries and overviews** NEW 55

Skills now present versioned summaries and Markdown overviews sourced from Web input or package metadata, including inheritance, restoration of package defaults, and review-bound presentation, without rewriting the original skill package.

— Explains behaviour but no named surface to act on v0.2.26

└──▷ **BREAKING ON UPGRADE**

! Upgrading applies Flyway migrations V120–V121; back up PostgreSQL and skill storage before upgrading. Schema rollback requires restoring the matching pre-upgrade backup.

WAS THIS USEFUL? 👍 👎

## AWS Amazon Bedrock AgentCore Feedback ⓘ

Runtime: Support for MCP protocol version 2026-07-28 (2026-10-13) BLOG ↗

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

Amazon Bedrock AgentCore Runtime now supports MCP version 2026-07-28, adding spec-aligned HTTP error responses, updated session handling, and a GET `/ping` health endpoint.

UPDATED SUPPORTED PLATFORM MCP servers on `microVMs` and `Instances` compute types now support Model Context Protocol version `2026-07-28`, with spec-aligned HTTP error responses and updated session handling.

MCP servers now support a recommended `GET /ping` health endpoint per the MCP 2026-07-28 spec.

WAS THIS USEFUL? 👍 👎

## DATA

◆ Vector DBs & RAG

# aegis-stack

Feedback ⓘ v0.19.1 NOTES ↗ RANK

A production-ready FastAPI platform with modular components and a built-in control plane.

> aegis-stack v0.19.1 adds a DNS management UI to Overseer for managing registrar records with audit trail, a new plugin creation flag, and removes two environment variables from the example config.

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

01 **--repository flag for aegis plugins create** NEW 75

Passing `--repository` to `aegis plugins create` sets the plugin's source repository at creation time, e.g. `aegis plugins create --repository https://github.com/example/my-plugin`, instead of configuring it after the fact.

Create a new plugin and point it at a remote repository in one step, rather than setting the repository after the fact.

```
$ aegis plugins create --repository
https://github.com/example/my-plugin
```

— Exact flag and runnable command given v0.19.1

02 **DOMAIN and ACME_EMAIL removed from .env.example** BREAKING 70

The `.env.example` no longer includes `DOMAIN` or `ACME_EMAIL`; projects with ingress and TLS that copied an older `.env` must remove those two lines or the app will fail to boot.

— Clear migration note naming exact env vars v0.19.1

03 **DNS record management in Overseer** NEW 62

From `Overseer > Deployments > DNS`, admins can add, edit in place, or delete DNS records on `BASE_DOMAIN` at the Porkbun registrar, with every change captured in a full audit trail.

— Names UI path and registrar but lacks audit trail detail v0.19.1

└──▷ BREAKING ON UPGRADE

! The `.env.example` no longer includes `DOMAIN` or `ACME_EMAIL`; remove those two lines from any `.env` copied before this release or the app will not boot on a project with ingress

and TLS.

---

WAS THIS USEFUL?  

# dosco GraphJin

Feedback ⓘ v3.21.19 NOTES ↗ RANK

One governed graph for AI agents — GraphQL + MCP over your databases, files, APIs, and code

> GraphJin's v3.21.19 release introduces a lead agent framework with persistent threads, resumable runs, and hot key/provider rotation, alongside cron- and metric-driven watch triggers, read-only OpenAPI content types, and cross-replica auth session sharing.

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

## 01 Lead agent with persistent threads and hot rotation NEW

75

GraphJin's agent system gains a lead agent that runs over stateless GraphJin workers holding skills, memories, and threads in the graph. It supports persistent threads and resumable runs with conversation state stored in the backing store, rotating the API key and switching the LLM provider without a server restart, agent profiles with home threads and per-thread unread state, naming the source of every figure in a lead answer by citing rows by their key, and scoring the lead in a data-accuracy evaluation loop.

— Rich mechanism described but no concrete command or config path v3.21.19

## 02 Cron and metric watch triggers NEW

65

GraphJin's watches now support `cron schedule triggers` to fire watch runs on a time-based schedule, metric watches on aggregates and agent routines, and watch routines posting answers directly into their agent's home thread.

— Names the watch mechanisms but no exact config syntax v3.21.19

THINNER COVERAGE BELOW

## 03 OpenAPI text/plain and text/html GET responses NEW

50

GraphJin now exposes `text/plain` and `text/html` read-only GET responses in its OpenAPI spec.

— Names the spec change but no endpoint or usage example v3.21.19

## 04 Cookie-based login for JWT-only servers NEW

45

GraphJin's web console now supports signing in to JWT-only servers using a cookie session.

— Clear UI capability but lacks configuration detail v3.21.19

## 05 Shared auth state across replicas IMPROVED

35

GraphJin now shares login and OAuth state across replicas in the auth store.

— States the behavior but not the storage mechanism v3.21.19

---

WAS THIS USEFUL? 👍 👎

# knowledge-rag

Feedback ⓘ v5.0.1 NOTES ↗ RANK

Local RAG MCP server for Claude Code — hybrid search (semantic + BM25), cross-encoder reranking, 13 MCP tools, 20 format parsers. Zero external servers, zero API keys.

> knowledge-rag v5.0.1 adds a cross-platform skills installer CLI and significantly reduces CUDA VRAM retention after indexing and search, alongside fixes to import side effects and GPU reranker settings.

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

## 01 Skills installer CLI NEW

85

A new `knowledge-rag skills install`, `knowledge-rag skills uninstall`, and `knowledge-rag skills status` command set installs, removes, or checks usage skills across Linux, macOS, and Windows, also runnable via `npx -y knowledge-rag skills install`. The installer detects coding assistants automatically, records a SHA-256 manifest of what it installed, and asks once per skills folder; uninstall removes only what the installer recorded.

Install usage skills for all detected coding assistants in one step — the installer detects them, records a SHA-256 manifest, and asks once per skills folder.

```
$ knowledge-rag skills install
```

Check installation status of usage skills across all detected coding assistants before or after an upgrade.

```
$ knowledge-rag skills status
```

Remove only what the installer recorded — useful for a clean uninstall without touching skills installed by other means.

```
$ knowledge-rag skills uninstall
```

— Named subcommands with mechanism (manifest, per-assistant detection) and runnable examples.

v5.0.1

## 02 CUDA VRAM released after indexing and search IMPROVED

65

CUDA VRAM now returns to the driver after indexing and after every search, measured on an RTX 3080 Ti as dropping from 1816 MiB to 456 MiB after indexing and from 4898 MiB to 466 MiB over 20 searches.

— Concrete before/after numbers but no config or flag to act on. v5.0.1

THINNER COVERAGE BELOW

03 **npx knowledge-rag forwards arguments** IMPROVED 55

`npx knowledge-rag` now forwards its arguments to the server instead of always starting a plain stdio server.

— Names the exact command behavior change, runnable as-is. v5.0.1

04 **Import no longer creates data folders** IMPROVED 50

Importing `mcp_server` no longer creates `data/`, `documents/`, or `models_cache/` folders as a side effect of import.

— Names exact folders fixed but lacks broader context. v5.0.1

05 **Reranker respects GPU embedding setting** IMPROVED 50

The cross-encoder reranker now respects the `models.embedding.gpu` setting.

— Names the config key but gives no further mechanism. v5.0.1

WAS THIS USEFUL? 👍 👎

# liliang-cn CortexDB

Feedback ⓘ v2.146.0 NOTES ↗ RANK

AI memory and a knowledge graph in one SQLite file. Pure Go: vectors, RAG, agent memory, RDF/SPARQL, Cypher, 80+ MCP tools.

> CortexDB's v2.146.0 release adds native media ingestion for images, audio and video, full version history for the knowledge graph, and a redesigned live view with a change stream and dedicated entity pages.

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

01 **Live view change stream and entity pages** IMPROVED 45

The live view has been redesigned with a change stream, a new Changes page, and an Entity page; an observed event can now name the specific nodes it lights by id.

— Names UI pages but no navigation path or config given v2.146.0

02 **Full knowledge-graph version history** NEW 43

CortexDB now tracks every version of the knowledge graph — triples, ontology, and which version is currently in effect — via a full graph version history.

— Explains what's tracked but no command or query shown v2.146.0

03 **Media ingestion as searchable knowledge** NEW 35

CortexDB can now ingest images, audio, and video as searchable knowledge using native media vectors, rather than being limited to text embeddings.

— Describes capability but no API, flag or format named v2.146.0

WAS THIS USEFUL? 👍 👎

# stashbase

Feedback ⓘ v2.18.11 NOTES ↗ RANK

stashbase is an open-source, local-first writing workspace where you write with Claude Code and Codex, drawing on your own sources.

> StashBase added anonymous download tracking for Gallery templates and moved its template repositories to a new organization.

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

01 **Anonymous download statistics for Gallery templates** NEW

42

StashBase now tracks anonymous download counts for Gallery templates, counting each completed copy once and retrying delayed reports without letting a retired template block others.

— Explains mechanism but no UI path or API named v2.18.11

02 **Template repositories moved to stashbase-org** IMPROVED

35

StashBase now links to the seven moved template repositories under the stashbase-org organization, with template IDs unchanged.

— Names the org and count but no actionable step v2.18.11

WAS THIS USEFUL? 👍 👎

# InfiniFlow RAGFlow

Feedback ⓘ v1.0.0 NOTES ↗ RANK

RAGFlow is a retrieval-augmented generation engine built on deep document understanding, combining chunk-level citations with agentic workflows.

> RAGFlow v1.0.0 expands its model and search ecosystem with new providers and adds operational controls for admins and ingestion pipelines, including audit logging, parsing-queue cancellation, and new database/OCR backends.

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

01 **Vastbase G100 database and MonkeyOCR PDF dispatch** NEW

60

Adds support for Vastbase G100 as the metadata database and `MonkeyOCR` PDF dispatch on the Go ingestion path.

– Names concrete products but no setup steps given v1.0.0

THINNER COVERAGE BELOW

02 **Parsing queue cancellation controls** NEW

55

Adds a one-click cancel control to remove all documents from the parsing queue and a cancel button in the UI to pause tasks that are currently running or queued.

– Describes UI controls but no exact navigation path v1.0.0

03 **Admin interface search, filtering and pagination** IMPROVED

50

Adds user search, status filtering, sorting, and CLI pagination to the admin interface.

– Lists concrete admin capabilities including CLI pagination v1.0.0

04 **Markdown and HTML remote images as media references** IMPROVED

40

Hands off Markdown and HTML remote images as media references during parsing.

– Explains behaviour but no format or API detail v1.0.0

05 **New model and web search providers** NEW

35

Adds Wally and Synap as new model providers, and `AnySearch` as a web search provider for chat agents.

– Names providers but no integration details Names the provider but lacks configuration details

v1.0.0

06 **Audit hooks for file and sync APIs** NEW 35

Adds audit hooks to file APIs and sync operations for access logging.

— States purpose but no endpoint or log format named v1.0.0

07 **Lifecycle events for document-parse tasks** NEW 35

Emits structured lifecycle events for document-parse tasks.

— Describes the feature but not the event schema v1.0.0

08 **User ownership for resources** NEW 25

Adds user ownership to resources.

— Bare description with no mechanism or scope v1.0.0

09 **Enterprise build-time overlay layer for web frontend** NEW 25

Adds an enterprise build-time overlay seam layer for the web frontend.

— Vague description with no usage path v1.0.0

WAS THIS USEFUL? 👍 👎

## infino

Feedback ⓘ v0.11.26 NOTES ↗ RANK

Embedded retrieval library built on Parquet. Fast, efficient, and scalable.

> Infino added a sorted top-k retrieval method and improved query planning to skip irrelevant superfiles for OR predicates, reducing I/O on filtered searches.

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

01 **Sorted top-k search by range-indexed column** NEW 60

Adds `sorted_search` to retrieve the first k rows ordered by a range-indexed column, enabling efficient top-k range queries.

– Names the exact API but no usage example or parameters v0.11.26

THINNER COVERAGE BELOW

02 **Superfile skipping for OR predicates** IMPROVED 45

Query OR conditions in WHERE clauses now skip non-matching superfiles, reducing I/O for filtered searches.

– Explains mechanism but no concrete numbers or usage v0.11.26

WAS THIS USEFUL? 👍 👎

## zvec

Feedback ⓘ v0.8.1 NOTES ↗

A lightweight, lightning-fast, in-process vector database

zvec adds grouped vector search to the C API, product quantization end-to-end, external vector sources for the flat index, and ARM NEON distance kernels.

└──▷ GET THIS VERSION ▾

```
$ git clone --branch v0.8.1 https://github.com/alibaba/zvec.git
# already have the repo? check out this version:
$ git checkout v0.8.1
```

Expose grouped vector search through the C API.

zvec's flat index now supports external vector sources with optional scan iteration.

zvec adds product quantization support end-to-end via the turbo engine.

zvec's turbo engine now includes ARM NEON one-to-many distance kernels, accelerating distance computation on ARM platforms.

NEW SUPPORTED PLATFORM RaBitQ on Linux/macOS/Android/iOS ARM64 and macOS x86_64

NEW SUPPORTED PLATFORM DiskANN on Intel macOS and Windows ARM64

WAS THIS USEFUL? 👍 👎

# redis-vl-python

Feedback ⓘ v0.31.0 NOTES ↗ RANK

Redis Vector Library. The AI-native Python client for Redis.

> redis-vl-python's v0.31.0 adds Python 3.15 support and overhauls how sync clients convert to async, while fixing long-standing bugs where zero-valued hybrid search parameters were silently dropped instead of sent to the server.

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

## 01 Zero-valued hybrid search parameters now honoured

BREAKING

80

`HybridQuery` now sends zero-valued combination parameters to the server instead of dropping them: `linear_alpha=0` and `rrf_constant=0` are now honoured, changing result ranking for affected queries, while `rrf_window=0` now returns a server error instead of being silently ignored. The MCP server's native hybrid search applies the same fix, now honouring `search.params.linear_text_weight: 0` and `rrf_constant: 0` instead of dropping them.

— Names exact fields across two surfaces with concrete before/after behaviour v0.31.0

## 02 Async client conversion preserves TLS, socket and retry settings

IMPROVED

70

Async paths that accept a sync Redis client now carry over TLS (Redis(ssl=True, ...)), unix-socket connections, and custom SSLConnection subclasses when converting, so these client types now work correctly in `SemanticCache`, `AsyncSearchIndex`, and other async paths. The converted async client also now preserves the sync client's retry policy, so it reconnects after a dropped connection instead of failing.

— Names mechanism and affected classes but no runnable example v0.31.0

## 03 Unsupported client types rejected at conversion time

BREAKING

65

Sentinel-managed clients, `RedisCluster` clients, and clients with sync-only settings such as OCSP validation now raise `ValueError` at conversion time and point to `redis_url`, instead of failing later with a `TypeError` on the first command.

— Clear before/after behaviour named, no code example v0.31.0

THINNER COVERAGE BELOW

## 04 Python 3.15 support NEW

55

redis-vl-python now supports Python 3.15 for all extras except where an upstream package's own Python bounds block it ( `voyageai` on 3.15, `sql-redis` on 3.14+); the `mcp` extra now requires `beartype>=0.23.1` .

— Names versions and extras but no usage detail `v0.31.0`

### └──▷ BREAKING ON UPGRADE

! `HybridQuery` previously ignored `linear_alpha=0` , `rrf_constant=0` , and `rrf_window=0` ; it now sends them, so queries using these zero values will rank differently or receive a server error for `rrf_window=0` .

! The MCP server's `search.params.linear_text_weight: 0` and `rrf_constant: 0` are now sent instead of dropped, changing result ranking for affected hybrid searches.

! Sentinel-managed clients, `RedisCluster` clients, and clients with sync-only settings (e.g. OCSP validation) now raise `ValueError` at conversion time instead of `TypeError` on the first command.

---

WAS THIS USEFUL?  

# oxideterm

Feedback ⓘ v2.3.1 NOTES ↗ RANK

AI-native workspace for local shells and remote machines.Zero Webview, zero OpenSSL, zero telemetry, and no app subscription.Written in Rust+GPUI

> OxideTerm v2.3.1 adds local Host Tools for machine inspection without SSH, brings tmux/Zellij/Herdr session management into its multiplexer panel, and shows SSH round-trip latency, alongside Session Manager, command palette, settings and rendering-performance improvements.

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

## 01 Multiplexer panel support for tmux, Zellij and Herdr NEW

80

The terminal multiplexer panel now supports tmux, Zellij and Herdr: you can expand a session to inspect tabs and panes (plus workspaces in Herdr), then attach, focus, create, split or close panes from within OxideTerm.

Attach to a remote tmux session and split a pane, all without leaving OxideTerm.

> 📍 In OxideTerm, open Host Tools for an SSH connection, open the Terminal Multiplexers panel, expand the tmux session, select a window, then click Split.

— Names three tools and the exact pane actions with steps v2.3.1

## 02 Host Tools for the local machine NEW

75

OxideTerm can now run Host Tools against the local machine — inspecting processes, ports, services, logs, filesystems, containers and packages without opening an SSH connection first. On Windows this extends to winget: inspection, upgrade and uninstall of packages by package ID.

Inspect open ports and running containers on the local machine without an SSH connection — useful for a quick local host audit.

> 📍 In OxideTerm, open Host Tools, select the local machine from the host selector, then choose Ports or Containers from the left panel.

— Names concrete inspection surfaces and winget actions with steps v2.3.1

03 **SSH round-trip latency display** NEW 70

The SSH terminal toolbar can now show request round-trip time for the active connection, with a default probe interval of 10 seconds. The interval is configurable, the display can be disabled, and timeouts are shown separately from valid readings.

— Gives default value and configurability but no exact setting path v2.3.1

04 **Faster terminal rendering via glyph reuse** IMPROVED 70

Terminal rendering now reuses unchanged glyphs and text layouts during streaming output and scrolling, reducing plain-text draw time from 1,108 to 773 microseconds and mixed-Unicode draw time from 1,133 to 824 microseconds on Windows x64.

— Precise mechanism and benchmark numbers but nothing to configure v2.3.1

05 **Session Manager batch tunnel starts** IMPROVED 65

Starting a tunnel from Session Manager no longer switches tabs; batch starts now authenticate once per saved host and start rules in sequence, with errors identifying affected tunnels by name.

Start multiple saved tunnels at once from Session Manager to authenticate once per host and avoid tab-switching interruptions.

> 📍 In OxideTerm, open Session Manager, multi-select the saved tunnel entries you want, then click Start — OxideTerm authenticates once per saved host and starts each rule in sequence.

— Describes behaviour change and gives usage steps v2.3.1

THINNER COVERAGE BELOW

06 **Settings and shortcut customization** IMPROVED 55

Settings gains new customization options: previous and next command-suggestion shortcuts are now configurable, on macOS the shortcuts for hiding the app and minimizing the window can be rebound in shortcut settings, and Appearance settings now let you choose between standard rounded corners and continuous curves.

— Names specific settings but no exact UI path or keys v2.3.1

07 **Command palette finds all saved connection types** IMPROVED 30

The command palette can now find and open all saved connection types.

— Thin one-line description with no detail on scope v2.3.1

WAS THIS USEFUL?  

◆ Document Parsing

# pdf-mcp

Feedback ⓘ v3.7.0 NOTES ↗ RANK

MCP server that lets Claude Code and other AI agents read and search large PDFs, one file or a whole folder: agentic RAG with hybrid semantic + keyword search, selective page reads, tables, images, OCR, chart data, and multi-column/CJK layouts.

> pdf-mcp's v3.7.0 release overhauls cache retention and configuration visibility, adds support for the newer sessionless MCP 2026-07-28 protocol, improves document title detection, and returns structured errors for unreadable PDFs alongside two breaking protocol-response changes.

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

01 **Configurable permanent cache retention** IMPROVED 86

Cache size and lifetime are now controlled via `PDF_MCP_CACHE_MAX_MB` (default 2048 MB, `0` for no cap) and optional `PDF_MCP_CACHE_TTL` ; cached documents persist as long as their PDF exists unchanged instead of expiring after 24 hours. The active configuration can be inspected at runtime through `server_info` , which now reports `config.cache_max_mb` and `config.cache_ttl_hours` (null when no TTL is set).

— Names exact env vars, defaults and inspection field v3.7.0

02 **MCP protocol 2026-07-28 support** NEW 70

pdf-mcp now supports MCP protocol 2026-07-28 (FastMCP 4, MCP SDK 2), allowing sessionless clients to connect over STDIO and HTTP without an `initialize` handshake or session IDs, while older clients including Claude Desktop continue to negotiate as before.

— Explains mechanism and backward compatibility clearly v3.7.0

03 **Smarter document title extraction** IMPROVED 67

`pdf_corpus_overview` and `pdf_corpus_search` now extract the real document title from the largest text at the top of page 1 when the embedded title is junk, skipping banners and section labels, and falling back to the filename as a last resort.

— Describes concrete heuristic and fallback behaviour v3.7.0

04 **Structured errors for unreadable PDFs** IMPROVED 62

Single-document tools now return a structured inline error with `error_code: pdf_unreadable` and a `hint` field when a damaged, truncated, or fake `.pdf` is opened, instead of failing the whole tool call with raw engine text.

— Names exact error code and response field v3.7.0

THINNER COVERAGE BELOW

05 **Protocol response behaviour changes** BREAKING 50

Reading a resource that does not exist now returns error `-32602` instead of `-32002`, and a cancelled tool call now gets no response at all instead of an error response.

— Names exact error codes changed but no migration guidance v3.7.0

└──▷ **BREAKING ON UPGRADE**

! Reading a resource that does not exist now returns error `-32602` instead of `-32002`.

! A cancelled tool call now gets no response at all instead of an error response.

WAS THIS USEFUL? 👍 👎

## kordoc

Feedback ⓘ v4.24.49 NOTES ↗ RANK

모두 파싱해버리겠다 — HWP·HWPX·PDF·Office 문서를 Markdown으로. 양식 자동 채우기와 신구 대조를 갖춘 CLI·MCP 서버 | Convert Korean documents (HWP, HWPX, PDF, Office) to Markdown — CLI and MCP server with form filling and diff

> kordoc added image format normalization and QR code reading from images, while cutting peak memory for image-heavy PDFs nearly in half via a PDFium-based decode path.

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

### 01 QR code reading from embedded images NEW 93

Read QR codes embedded in images of any format with the `--qr-codes` CLI flag or `qrCodes` API option (protocol 2 and both SDKs); results appear in the image block's `qr` field and on the line after the Markdown image placeholder. Off by default; loads optional dependency `zxing-wasm` only when enabled, with no network access.

Extract QR code values from every image in a scanned PDF alongside the normal Markdown output.

```
$ npx kordoc 스캔본.pdf --ocr --qr-codes -o 스캔본.md
```

— Flag, API option, output location, dependency, and default all named v4.24.48

### 02 Normalize extracted images to PNG or WebP NEW 83

Pass `--image-format png` or `--image-format webp` (lossless) on the CLI, or set `imageFormat: "png" | "webp"` in the API, to normalize all extracted images to a single format in one pass — covers BMP, TIFF, and files whose name and content type did not match.

Convert a scanned PDF to Markdown while forcing all embedded images to lossless WebP for consistent downstream processing.

```
$ npx kordoc 스캔본.pdf --image-format webp -o 스캔본.md
```

— Names exact flag, values, and API key with coverage detail v4.24.49

### 03 PDFium decoding cuts PDF image memory IMPROVED 73

PDF images wider than 128px now decode via PDFium instead of pdfjs, cutting peak memory from 1,761 MB to 868 MB on image-heavy documents; block structure, Markdown output, image numbers and counts are unchanged.

— Concrete before/after numbers but no CLI action required v4.24.48

---

WAS THIS USEFUL? 👍 👎

## EVALUATE

◆ AI Observability & Evals

# LangChain LangSmith

Feedback ⓘ updated · Nov 7, 2026

NOTES ↗ RANK

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

> LangSmith's biggest changes this window are to the LLM Gateway — IAM role authentication for Amazon Bedrock and detailed gateway metadata surfaced in traces — alongside thread-level annotation queue automation (which now requires SmithDB queries), IPv6 support, three new GPT-5.6 chat models, and sandbox usage-history and WebSocket fixes.

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

01 **LLM Gateway metadata surfaced in traces** NEW 75

LLM Gateway responses now include an `X-LangSmith-Gateway-Metadata` response header, attachable to traces as `ls_gateway_info` metadata, describing gateway routing, limits, guard activity, and upstream errors. Trace details now surface LLM Gateway outcomes, policy limits, selected models, and fallback attempts in a dedicated Gateway section, and errored runs blocked by the gateway show a Blocked badge in the trace tree.

— Names exact header and metadata key but no setup steps snapshot-20261107

02 **Thread-level annotation queue automation** BREAKING 70

Thread-grouped automations can now add one annotation queue item per matched conversation so reviewers can inspect the whole thread instead of only one run. Thread webhook and annotation-queue automations now show thread filters instead of run filters and no longer default to `Is Trace = true`, and both automation types now require SmithDB queries, meaning existing automations using other query types will need to be migrated.

— Specifies filter and query changes but no exact migration steps snapshot-20261107

03 **Turns view deprecated in thread detail** DEPRECATED 70

The Turns view in thread detail is deprecated and will be removed from LangSmith Cloud on October 31, 2026 and from self-hosted in v17; use the Messages or Details view instead.

— Gives exact removal dates and replacement views snapshot-20261107

04 **CSV export for colon-containing dataset keys** IMPROVED 65

Datasets whose input, output, or metadata keys contain a colon — including `source_run_feedback:<key>` metadata keys — can now be exported to CSV.

— Names exact metadata key pattern enabling export `snapshot-20261107`

05 **IAM role authentication for Bedrock gateway** NEW 60

The LLM Gateway can now authenticate to Amazon Bedrock using an assumed IAM role or AWS access keys instead of requiring a static Bedrock API key, with existing Bedrock API keys taking precedence when multiple credential types are configured.

— Explains credential precedence but no exact config key `snapshot-20261107`

06 **WebSocket query strings preserved in sandbox URLs** IMPROVED 60

WebSocket connections through a sandbox service URL now reach your service with their query string intact, allowing dev servers that authenticate WebSocket connections with a query parameter (such as Vite's HMR token) to connect.

— Explains behavior with a concrete example (Vite HMR token) `snapshot-20261107`

THINNER COVERAGE BELOW

07 **Sandbox CPU, memory, network usage history** NEW 45

Sandboxes now retain recent CPU, memory, and network usage history, making it easier to diagnose performance bottlenecks.

— States metrics retained but no retention window or UI path `snapshot-20261107`

08 **Full-list filtering on API Keys page** IMPROVED 45

The API Keys page now filters keys across the full paginated list by key, description, and workspace.

— Names filter fields but no UI path or limits `snapshot-20261107`

09 **GPT-5.6 Sol, Terra, Luna in LangSmith Chat** NEW 40

LangSmith Chat now offers GPT-5.6 Sol, GPT-5.6 Terra, and GPT-5.6 Luna alongside the existing OpenAI models.

— Names three new models but no usage detail `snapshot-20261107`

10 **Tag visibility for tag-on-create permission** NEW 40

Users with the tag-on-create permission for datasets, projects, or prompts can now see and apply tags when creating those resources.

— Names permission and resource types but no UI detail `snapshot-20261107`

11 **IPv6 connectivity for LangSmith services** IMPROVED 25

LangSmith services now accept both IPv4 and IPv6 connections.

— Bare statement with no scope or mechanism snapshot-20261107

### └──▷ BREAKING ON UPGRADE

! Thread annotation-queue and webhook automations now require SmithDB queries; existing automations using other query types will need to be migrated.

! The Turns view in thread detail is deprecated and will be removed from LangSmith Cloud on October 31, 2026 and from self-hosted in v17; use the Messages or Details view instead.

---

WAS THIS USEFUL?  

# UK AI Security Institute Inspect AI

Feedback ⓘ

0.3.288 CODE ↗ RANK

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

> Inspect AI 0.3.288 adds visibility and control for multi-agent runs — new transcript events for agent swarms, a CLI command to list named agents, and per-agent sandbox user isolation — alongside several breaking changes including removal of the legacy web browser tools.

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

01 **google-genai 3.0 version requirement** BREAKING 75

Environments using `google-genai` 2.x now fail the `validate_google_client` version check — install `google-genai>=3.0.0` to continue.

— Exact package, check name and fix version given 0.3.288

02 **Removal of web_browser() tools** BREAKING 75

Calls to web_browser() or its web_browser_tools() alias are removed — replace with web_search() for retrieval or computer() / sandbox MCP tools for browser interaction.

— Names removed functions and exact replacements 0.3.288

03 **`inspect ctl sample agent` CLI subcommand** NEW 65

A new `inspect ctl sample agent` subcommand lists a sample's named agents.

— Runnable command named but usage/output undetailed 0.3.288

04 **Default Docker sandbox image changed** BREAKING 65

Evaluations using the automatically generated Docker sandbox must now supply an explicit image if they need the `nonroot` account, `pipx`, or Playwright browser dependencies, since the default image is now the official Python image.

— Explains before/after but not the new image's name 0.3.288

THINNER COVERAGE BELOW

05 **Runtime continuation of agent loops** NEW 55

A runtime can now keep a react() or deepagent() loop running after the agent answers, continuing the loop rather than stopping.

— Names the functions but not how to configure continuation 0.3.288

06 **Per-agent sandbox user isolation** NEW 50

Custom tools can now run inside a sandbox identity scope as that scope's own OS user, isolating each agent's sandbox tools by identity.

— Mechanism described but no config key or flag given 0.3.288

07 **Agent swarm transcript events** NEW 40

Experimental transcript events capturing agent-to-agent communication and agent swarm evidence are now added to the eval log.

— No flag or command shown to enable or inspect it 0.3.288

### └──▷ BREAKING ON UPGRADE

! Environments using `google-genai` 2.x now fail the `validate_google_client` version check — install `google-genai>=3.0.0`.

! Evaluations using the automatically generated Docker sandbox must supply an explicit image if they need the `nonroot` account, `pipx`, or Playwright browser dependencies (the default image is now the official Python image).

! Calls to web_browser() or its web_browser_tools() alias are removed — replace with web_search() for retrieval or computer() / sandbox MCP tools for browser interaction.

WAS THIS USEFUL?  

## PromptLayer

Feedback ⓘ November 06, 2026 NOTES ↗ RANK

PromptLayer is a platform for managing, evaluating, and observing prompts and LLM applications.

> PromptLayer updated default models and token limits across its prompt and playground experiences for Anthropic, Bedrock, OpenRouter, OpenAI, and Azure OpenAI.

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

01 **New default models and token limit in playground** IMPROVED

55

New prompts and playground chats now default to `Claude Haiku 5.5` for Anthropic, Amazon Bedrock, and OpenRouter. OpenAI and Azure OpenAI playground defaults now use `GPT-6 Luna`. Chat models in the playground now default to a `128,000` max token limit.

— Names exact default models and token limit, but only a config default change.

November 06, 2026

WAS THIS USEFUL? 👍 👎

## Braintrust

Feedback ⓘ updated · Nov 7, 2026 NOTES ↗

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

Braintrust adds trace continuation via `--parent` in `bt trace run`, ephemeral Terraform resources, Patterns problem detection, thread-level English translation, and 2FA visibility for org owners.

Continue a trace from a coding-agent run by passing `--parent` to `bt trace run` with the value returned by an SDK span's export(), attaching the session below an existing span instead of starting a new trace.

Calling a Braintrust MCP server tool now works without specifying a telemetry argument — `bt mcp run` no longer requires it.

Suppress async-context span errors in the Python SDK by setting the environment variable `BRAINTRUST_OTEL_COMPAT=true`.

Manage saved views for Patterns as code in the Terraform provider by setting `view_type` to a pattern value in the `braintrust_view` resource.

Create a service token without storing its key in Terraform state or plan files using the new ephemeral `braintrust_service_token` resource.

+8 more

NEW INTEGRATION Firebase Genkit tracing (Go SDK)

NEW INTEGRATION MCP server tracing for fastmcp and mcp-handler (TypeScript SDK)

UPDATED INTEGRATION TypeScript SDK tracing support for newer SDK versions

UPDATED INTEGRATION Anthropic Java SDK 3.x tracing support

NEW INTEGRATION Temporal activity tracing (Python SDK)

UPDATED INTEGRATION Cursor subagent and LiveKit Agents tracing (TypeScript SDK)

└──▷ **BREAKING ON UPGRADE**

! The Terraform provider now requires Terraform or OpenTofu 1.11.1 or later.

! `braintrust_ai_secret` now requires `secret_version` alongside `secret`, and `braintrust_env_var` now requires `value_version` alongside `value`. Changing only the value no longer updates it; you must increment the version. References to `secret` or `value` return `null`.

! The `braintrust_org_automation` resource is removed from the Terraform provider because the Braintrust API no longer supports it. Delete its blocks from your configuration, then run `terraform state rm` on each address before running `terraform plan`.

WAS THIS USEFUL?  

## AI MODELS

◆ Open-Weight Models

# Alibaba Qwen Models

Feedback ⓘ updated · Nov 7, 2026

NOTES ↗ RANK

Qwen is Alibaba's language and multimodal model family, including coding, vision, image, speech, and reasoning models.

> Qwen Model Studio added more than 25 new models in this window, spanning real-time speech, text-to-speech, ASR, omni-modal and video generation, image generation and editing, structured decision-making, and world simulation.

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

## 01 Real-time duplex speech models NEW

85

Qwen added `qwen-audio-3.1-realtime-flash` and `qwen-audio-3.1-realtime-plus`, real-time duplex speech models with a 262,144-token context window, function calling, web search, and voice cloning. `qwen-audio-3.1-realtime-flash` also supports text and audio input/output, while `qwen-audio-3.1-realtime-plus` uses the same integration protocol as 3.0 Plus and adds eight new system voices.

— Names both model IDs and their distinct capabilities. snapshot-20261107

## 02 Text-to-speech models NEW

85

Qwen added `qwen-audio-3.1-tts-next`, `qwen-audio-3.1-tts-plus`, and `qwen-audio-3.1-tts-flash`. `qwen-audio-3.1-tts-next` unifies speech, sound effects, ambient sounds, and soundscapes in a single pass from text, timestamps, and reference audio, with multilingual TTS, multi-speaker dialogue, podcasts, and cinematic audio. `qwen-audio-3.1-tts-plus` is a professional TTS model supporting multiple languages and dialects, streaming synthesis, instruction following, fine-grained tags for emotion, tone, persona, speaking rate, and volume, and noise-robust voice cloning. `qwen-audio-3.1-tts-flash` provides streaming speech synthesis for real-time interactions such as voice assistants and customer service.

— All three model IDs and their specific features are named. snapshot-20261107

## 03 ASR models for transcription and translation NEW

85

Qwen added `qwen-audio-3.1-asr-flash-message`, `qwen-audio-3.1-asr-flash-streaming`, `qwen-audio-3.1-asr-flash`, and `qwen-audio-3.1-asr-flash-filetrans`. `-message` is a generative speech recognition model for voice messages and enterprise scenarios with incremental streaming transcription, P0/P1 hotwords, text normalization, multi-speaker stability, automatic language detection, and Chinese dialect transcription. `-streaming` is a low-latency end-to-end real-time ASR

model for conferencing and live subtitles with multilingual support, hot word enhancement, punctuation prediction, and controllable ASR/AST output for multiple dialects. The base `-flash` model is a large-scale multilingual ASR model with Chinese dialect recognition, classical Chinese poetry optimization, multi-person dialogue role-based transcription, and controllable ASR/AST dialect output. `-filetrans` is an offline ASR model for meeting transcription and call analysis with speaker separation, high-precision transcription, hot word enhancement, and stable performance in noisy environments.

— Four model IDs each with named scenario-specific capabilities. snapshot-20261107

04 **Image generation and editing models** NEW 75

Qwen added `qwen-image-2.1-pro`, a 7B-parameter image model unifying text-to-image generation and editing with native transparent image support, up to 10 reference images, local edits, and subject preservation, and `qwen-image-2.1-turbo`, an open-source image generation model with native transparency, unified generation and editing, multi-reference editing, and local control.

— Named model IDs with parameter count and capability specifics. snapshot-20261107

05 **DeepSeek V4.1-Flash iteration with self-feedback RL** IMPROVED 75

Qwen added `deepseek-v4.1-flash-1017`, an iteration of DeepSeek V4.1-Flash trained with self-feedback reinforcement learning, retaining the 552B-parameter MoE architecture, native multimodal visual understanding, a 1M-token context window, and a 384K-token maximum output, while improving benchmark scores on Terminal-Bench 4.0, DeepSWE v1.1, and ProgramBench.

— Names model ID, parameters, context limits, and benchmarks. snapshot-20261107

06 **Omni-modal and real-time multimodal models** NEW 70

Qwen added `qwen3.8-omni-flash`, an omni-modal model supporting text, image, audio, and video inputs with thinking and non-thinking modes, function calling, web search, context caching, and agent applications, and `qwen3.8-omni-flash-realtime`, a real-time multimodal model supporting audio and video interaction with multichannel audio, video aggregation, remote MCP tools, and WebSocket, WebRTC, and AOQ access.

— Names both models and their access protocols. snapshot-20261107

07 **Structured decision model for business routing** NEW 65

Qwen added `decision-model-preview`, a structured decision model for high-frequency business decisions that runs classification, yes/no decisions, and scoring in parallel from text or business context, returning probability distributions and confidence scores for ticket routing, content moderation, agent routing, and result validation.

— Describes mechanism and use cases but no interface details. snapshot-20261107

08 **PixVerse video generation models** NEW 65

Qwen added `pixverse/pixverse-v6-t2v`, `pixverse/pixverse-v6-it2v`, `pixverse/pixverse-v6-kf2v`, and `pixverse/pixverse-v6-r2v` video generation models supporting up to 15-second videos with synchronized audio, multilingual text, and camera movement control including push, pull, pan, truck, and tracking.

— Names four model IDs and shared capability limits. snapshot-20261107

THINNER COVERAGE BELOW

09 **World-simulation models for interactive scenes** NEW 50

Qwen added `happyoyster-1.0-adventure`, `happyoyster-1.0-directing`, and `happyoyster-1.0-acting` world-simulation models that generate real-time interactive open-world scenes and characters from multimodal text/image input.

— Names three model IDs but little mechanism detail. snapshot-20261107

WAS THIS USEFUL? 👍 👎

## Liquid AI LFM3-1.7B (Liquid AI) Feedback ⓘ

new snapshot · Nov 7, 2026 MODEL ↗

LFM3-1.7B is a 1.7-billion-parameter language model from Liquid AI, designed for efficient on-device and edge deployment rather than as a security or developer tool.

New snapshot 20261106 listed on OpenRouter (liquid/lfm-3-1.7b), first seen 7 Nov 2026.

WAS THIS USEFUL? 👍 👎

## ConvAI Innovations Laya

Feedback ⓘ v0.4.27 NOTES ↗

RANK

Non-autoregressive System 1 decision engine. Typed choice, score and yes/no decisions over any text in a single forward pass, in 100+ languages, with a router that picks the right checkpoint per request.

> Laya's TypeScript SDK gained support for HTTP batch predictions.

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

01 **HTTP batch predictions in TypeScript SDK** NEW 30

Laya's TypeScript SDK now supports HTTP batch predictions.

— No mechanism, limits, or usage details given v0.4.27

WAS THIS USEFUL? 👍 👎

## DeepSeek V3

Feedback ⓘ price increase · Nov 6, 2026 MODEL ↗

DeepSeek V3 is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-chat): $0.2574 → $0.32 in, $1.0287 → $0.89 out per million tokens (+24%), first seen 6 Nov 2026.

WAS THIS USEFUL? 👍 👎

## DeepSeek V4 Flash Vision Exp

Feedback ⓘ

price increase · Nov 6, 2026 MODEL ↗

DeepSeek V4 Flash Vision Exp is a multimodal model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4-flash-vision-exp): $0.2156 → $0.44 in, $0.6468 → $1.32 out per million tokens (+104%), first seen 6 Nov 2026.

WAS THIS USEFUL? 👍 👎

## Meta Llama 3.3 70B Instruct

Feedback ⓘ

price increase · Nov 6, 2026 MODEL ↗

Llama 3.3 70B Instruct is a chat model from Meta.

Price increase on OpenRouter (meta-llama/llama-3.3-70b-instruct): $0.10 → $0.59 in, $0.32 → $0.79 out per million tokens (+490%), first seen 6 Nov 2026.

WAS THIS USEFUL? 👍 👎

## Alibaba Qwen3 Next 80B A3B Instruct

Feedback ⓘ

price cut · Nov 6, 2026 MODEL ↗

Qwen3 Next 80B A3B Instruct is a chat model from Alibaba.

Price cut on OpenRouter (qwen/qwen3-next-80b-a3b-instruct): $0.10 → $0.09 in, $1.10 → $1.10 out per million tokens (-10%), first seen 6 Nov 2026.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · November 7, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue