THE AI TOOLCHAIN

NO. 098

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED NOVEMBER 22, 2026 · EVERY DAY

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

▸ // HOW THIS ISSUE IS MADE SHOW

VIEW

ISSUE VIEW

scroll



SCROLL | EXPAND IN PLACE

INFORMATION DENSITY

everything



BRIEFED | **EVERYTHING**

Every tool in the issue, in full.

all 80 changes

Do you prefer **scroll**?

▸ // CONTENTS SHOW 24 tools

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

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

Three unrelated things stand out today: Kiro can now run the IDE against remote SSH, Dev Container and cloud workspaces; Anthropic added

vim ? help

customer-linked billing and automatic prompt caching on Bedrock; Together AI made cluster storage reads faster with a free NVMe cache.

BUILD

Use a full coding workspace without cloning it onto your laptop

Kiro's desktop IDE can now attach to SSH hosts, Dev Containers and cloud sessions, with VPC connections for cloud sessions. That moves agentic coding closer to the actual build environment, so teams stop reproducing local dependency drift before the agent can work.

Kiro

OPERATE

Charge and optimize model usage closer to the real customer or workload

Anthropic added customer-linked billing and automatic prompt caching on Bedrock, Pinecone now exports billing data as FOCUS v1.4 CSV, and OpenHands can route each Agent Canvas task by cost and performance. The common win is less spreadsheet reconciliation and fewer hard-coded model choices when usage needs to map back to teams, customers or tasks.

Anthropic · Pinecone · OpenHands

DEPLOY

Keep model and agent infrastructure closer to your own network boundary

Zeroshot can now self-host its target on Kubernetes through a versioned Helm chart, while Arcade can route one gateway across multiple accounts per provider. That gives platform teams cleaner deployment and account separation instead of multiplying one-off service instances or credentials.

Zeroshot · Arcade

Does **GPT-5.5** read better? 👍 👎

OFFENSIVE SECURITY

◆ Exploitation & C2

xalgorix

Feedback ⓘ v4.6.203 NOTES ↗ RANK

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

> xalgorix shipped a major system prompt size reduction via deferred phase content, added per-scan PID namespace isolation for process security, and fixed scan resume notification and startup queue handling.

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

01 **Phase playbook deferral shrinks agent system prompt**

Not checked yet IMPROVED 82

Enabling `XALGORIX_PHASE_PLAYBOOKS` in the dashboard's Runtime settings cuts the main agent system prompt from about 114 KB to 59 KB (48%) by deferring each large phase body (phases 1–19) and injecting it only while tasks belonging to that phase are active or upcoming.

— Explains mechanism, exact size numbers, and config key. v4.6.203

02 **Per-scan PID namespace isolation** Not checked yet NEW 80

Each scan's processes are isolated in its own PID namespace so that `ps`, `kill`, `pkill`, and `killall` inside a scan can only reach that scan's own processes, not the host or other scans. Isolation can be controlled via the `XALGORIX_PROCESS_ISOLATION=false` environment variable or the dashboard's Runtime settings.

— Named commands, env var, and toggle location given. v4.6.202

THINNER COVERAGE BELOW

03 **Scan resume and queue restart fixes** Not checked yet IMPROVED 42

A resumed scan queue now sends a Scan Resumed notification instead of a duplicate Scan Started notification, including when resuming a paused scan. Startup auto-resume now resumes every interrupted queue independently rather than skipping all queues when any scan is already running.

— Describes behavior change but no config surface named. v4.6.202

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

Amazon Kiro

Feedback ⓘ changelog · Nov 22, 2026 CHANGELOG ↗

RANK

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

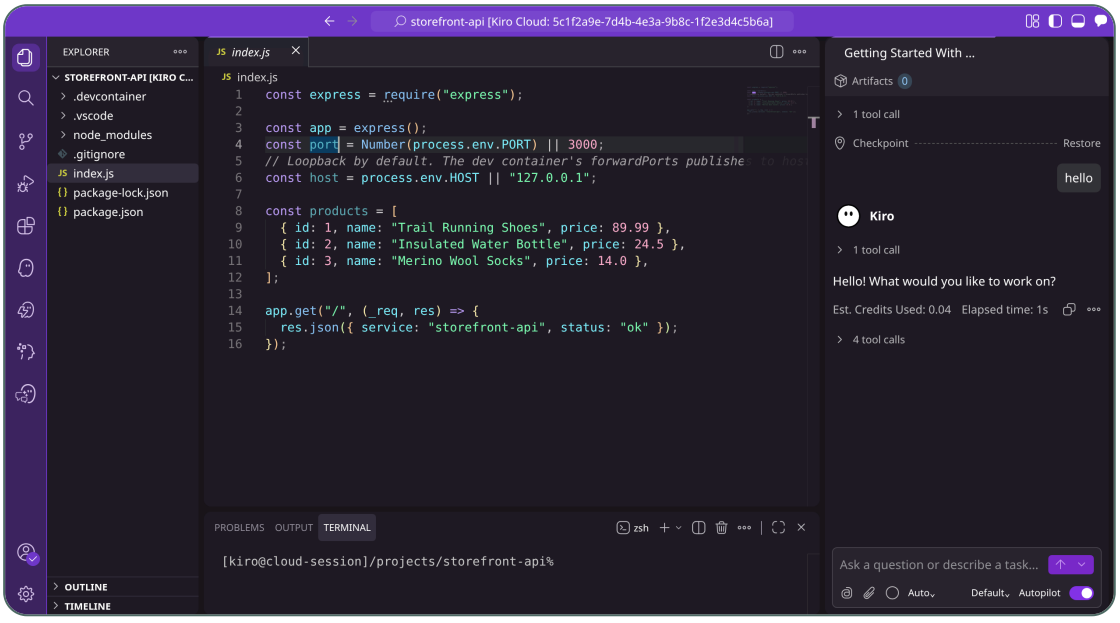
> Kiro's largest window addition is full remote workspace support — SSH, Dev Containers, and cloud-session IDE access from the desktop IDE — alongside a new VS Code extension, VPC connections for cloud sessions, and conductor mode, a multi-agent coordination layer in Kiro Crew; it also tightened defaults around telemetry, write safety, and settings organization.

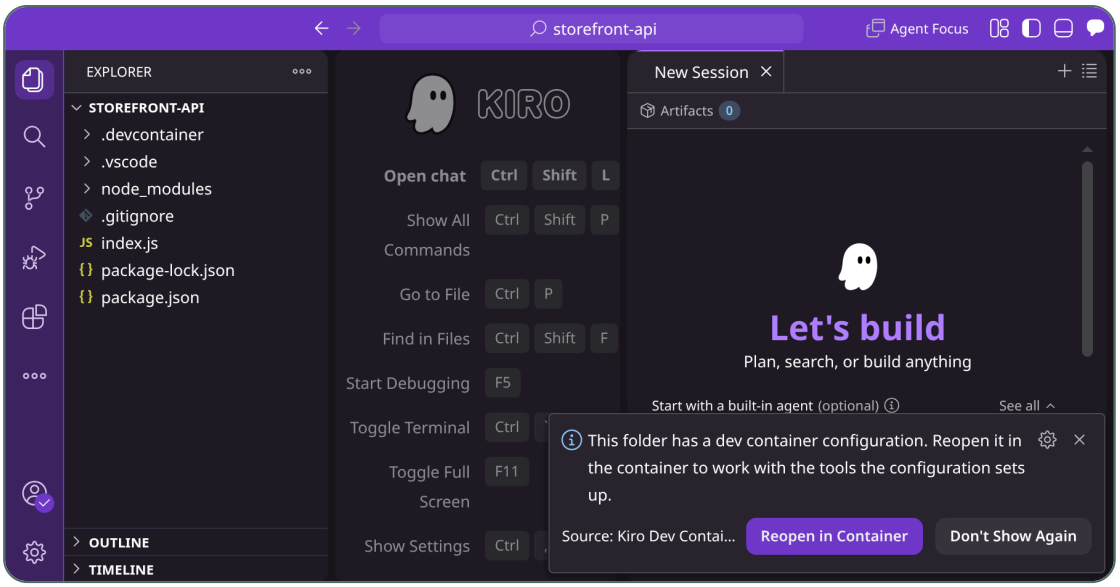
└──▷ WHAT SHIPPED · 9 FEATURES most completely described first ? what's the number?

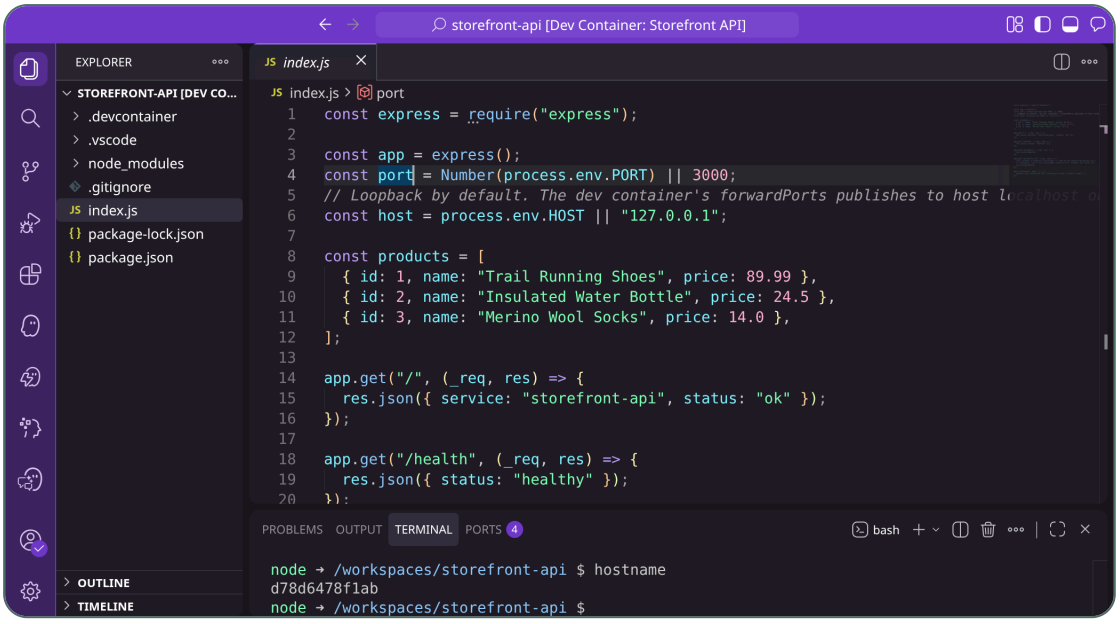
01 Remote workspace connections: SSH, Dev Containers, cloud sessions

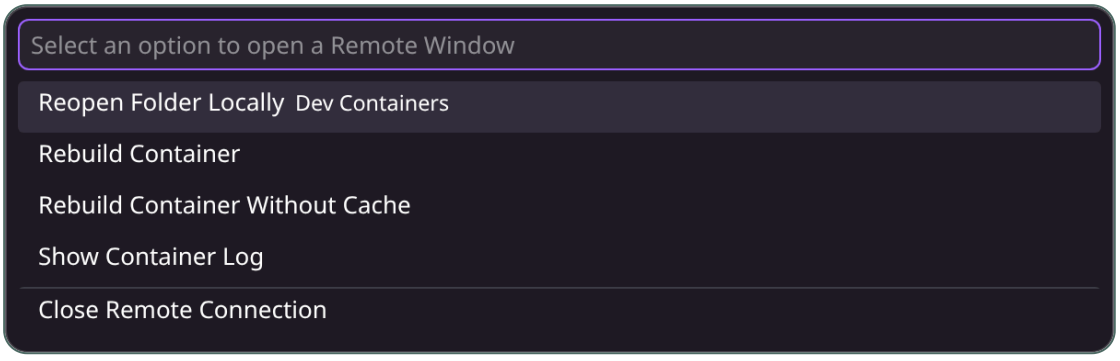
Not checked yet NEW 93

Connect the full Kiro IDE to a Linux SSH host (x86-64 or ARM64, glibc 2.28+, 4 GiB free in home directory) by running `Kiro: Add New SSH Host…` then `Kiro: Connect to Host…`, with hosts also browsable under `Remote environments > SSH Hosts`; open a repository's `.devcontainer/devcontainer.json` with `Reopen in Container` or the command palette's `Dev Containers: Reopen in Container`, and rebuild with `Rebuild Container`; and attach the editor, file tree, and terminal directly to a running cloud session with `Open in IDE` from the Sessions sidebar or Agent Focus Mode.









Connect to a Linux build box or GPU machine already on your network without installing any extension.

📍 1. Open the command palette and run `Kiro: Add New SSH Host…`, enter your host details, and save. 2. Run `Kiro: Connect to Host…`, select the saved host, and open a folder on it. 3. The window title shows the connected host; browse files and run terminals on the remote machine directly.

Reproduce the exact environment your CI uses by opening your repository's dev container locally with Podman.

📍 1. Ensure your repository contains `.devcontainer/devcontainer.json` and that Podman is running. 2. Open the command palette and run `Dev Containers: Reopen in Container`. 3. Kiro builds or attaches to the container; to pick up definition changes later, choose `Rebuild Container` from the remote menu.

Run a heavy build on Kiro-managed cloud compute so it does not stall your laptop, while keeping your normal editor experience.

📍 1. Open the Sessions sidebar or Agent Focus Mode in Kiro IDE. 2. Right-click an existing cloud session and choose `Open in IDE`. 3. A separate remote window opens; the file tree, editor, and integrated terminal all operate on the cloud session's filesystem.

— Names exact commands, requirements and three distinct connection paths

changelog-20261122-0c2612bb launch-20261120-72d17a20

---

02 **Conductor mode: multi-agent coordination in Kiro Crew**

Not checked yet NEW 90

Kiro Crew's conductor mode (replacing the earlier Ultra Mode toggle, now on by default in every standard build with the Settings > Developer > Feature Previews switch removed) has one conductor agent break a goal into parallel tasks, dispatch them to worker agents, and track results on a shared board. It validates task completion by calling `accept_eval`, which checks evidence (a pull request's checks, a required file, a pending approval) against each task's agreed completion condition, closing a task as rejected after three failures. State is stored in an append-only crew log with a work ledger (tasks and results) and session ledger (goal, next step, rejected approaches) so the conductor can resume after a dropped connection; patrol checks skip a model turn when the ledger is unchanged and move up when a worker reports done, blocked, or question; long idle sessions restart their next turn in a fresh model session loaded with the ledger snapshot; and the default conductor agent has no file-editing tool, delegating all changes to worker agents.

— Deep mechanism and named internals, but no runnable command given

launch-20261120-c7b56690 changelog-20261122-0c2612bb

---

03 **MCP server sign-in: organization IdP and CLI** Not checked yet NEW 83

Sign in once to organization MCP servers through your identity provider by adding the provider to your user MCP configuration and selecting `Sign in` on the `Organization sign-in` row in the MCP Servers view; in a V3 CLI session run `kiro-cli mcp idp setup` with your provider's issuer and client ID, then open `/mcp` and press Enter, using `--`

`oauth-client-id <client-id>` on `kiro-cli` for servers with a Kiro-issued client ID; and configure a custom login portal for `kiro-cli login` to control where browser sign-in starts.

— Names exact CLI commands and flags for three sign-in paths changelog-20261122-0c2612bb

04 **Kiro for VS Code extension** Not checked yet NEW 75

A new Kiro extension for VS Code reads the same `.kiro` folder as Kiro IDE so specs, steering files, skills, hooks, powers, and MCP servers work unchanged; start a vibe or spec session from Kiro's hub in the side bar, reference context with `#` (file, folder, open files, diagnostics, spec, steering file, terminal output), run `Kiro: Ask Kiro` from the Command Palette on a code selection, and Kiro respects VS Code Workspace Trust by not starting workspace-defined MCP servers in untrusted folders.

Ask Kiro about a specific code selection without leaving VS Code — useful for quick inline explanations or security reviews of a function.

📍 1. Select the code you want to ask about in the editor. 2. Open the Command Palette (Ctrl+Shift+P / Cmd+Shift+P). 3. Run `Kiro: Ask Kiro`.

Ground a Kiro session in specific project context — such as a steering file or a spec — to get answers that respect your team's conventions.

📍 1. Open the Kiro tab in the VS Code side bar. 2. In the chat input, type # and select a steering file, spec, or terminal output from the picker. 3. Type your question and start the session.

— Names commands and context syntax, but no config file surfaced launch-20261120-ddc6bbc9

05 **GitLab skill and powers installs** Not checked yet NEW 60

Custom powers and skills can be installed from a public GitLab URL the same way as from GitHub; add a self-hosted GitLab instance's hostname to the `kiroAgent.powers.trustedGitLabHosts` setting to allow it.

Allow skills and powers to be installed from a self-hosted GitLab instance by trusting its hostname in Kiro settings.

📍 In Kiro settings, add your self-hosted GitLab hostname to `kiroAgent.powers.trustedGitLabHosts`, then paste the GitLab URL in the powers/skills installer to install.

— Named config key and install path, but thin overall changelog-20261122-0c2612bb

THINNER COVERAGE BELOW

06 **Session continuity: reopen, fork, quick reply** Not checked yet NEW

55

Reopen the last closed session with Cmd+Shift+T (Ctrl+Shift+T on Windows/Linux, Alt+Shift+T in a browser tab), continue in a copy of a conversation with `/fork <prompt>`, and reply to a session awaiting input by hovering over it or pressing the right arrow key for a quick-reply card, togglable via Quick Reply Card on Hover under Settings > Chat > Sessions.

– Exact shortcuts and a slash command named, no deeper mechanism

changelog-20261122-0c2612bb

07 **Permission Request Hook** Not checked yet NEW

45

A new Permission Request Hook, created in the Hook editor, runs a command before Kiro asks for tool-call approval, letting a script answer routine prompts in unattended sessions.

– Describes purpose and location but no example or config shown

changelog-20261122-0c2612bb

08 **Composer and transcript chat settings** Not checked yet NEW

45

Rewrite a message before sending with Optimize Before Sending under Settings > Chat > Composer, and set the reply language with Preferred Reply Language under Settings > Chat > Transcript.

– UI paths named but behavior described only briefly changelog-20261122-0c2612bb

09 **Files tab default view** Not checked yet IMPROVED

35

The Files tab's new default view shows files the current chat created or changed, newest first, with a Browse option to switch to the full project folder.

– Simple UI change with no config or command surface changelog-20261122-0c2612bb

└──▷ ALSO FROM THESE RELEASES


0:00 / 0:24


0:00 / 0:11



┗──▷ BREAKING ON UPGRADE

! Ultra Mode is renamed to conductor mode and removed from Settings > Developer > Feature Previews — it is now on by default in every standard build with no switch to enable it.

! The Releases page is removed; release notes now live in Settings > About.

! MCP Servers moved from Settings > Connections to their own tab under Customize, beside Skills.

! Usage data (`telemetry.enabled`) is now on by default in public builds; opt out with `telemetry.enabled: false` , `KIROCREW_TELEMETRY=0` , or `DO_NOT_TRACK=1` .

! `hooks.batch_write_guard` now refuses a write sent beside a read when it quotes an ID or value the agent has not seen; it is on by default — set it to `false` to turn it off.

! On Linux, external edits to `config.json` , `config.local.json` , `denied_commands.json` , or `computer_use.json` that Kiro Crew does not recognize are held for review before being applied.

WAS THIS USEFUL? 👍 👎

yetone magpie

Feedback ⓘ v0.1.4676 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 latest releases expand agent detection to Mistral's Vibe Desktop app, add Nix/Guix-aware path handling for autostart entries, support Tailscale MagicDNS hostnames for remote connections, fix Claude Code turn continuation after cancellation, add quota warm-up for Antigravity's agy CLI, and let users hide individual Library tabs.

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

01 **Nix/Guix store-aware executable resolution** Not checked yet

IMPROVED 85

Reads the `MAGPIE_EXECUTABLE` environment variable to identify the Nix or Guix wrapper binary and uses it instead of os.Executable() when registering autostart entries, desktop entries, and PATH links. Binaries living under `/nix/store/` or `/gnu/store/` are mapped to their stable profile link so registered entries survive package upgrades and garbage collection.

— Mechanism and env var fully named, concrete before/after v0.1.4674

02 **Agy CLI quota warm-up for Antigravity accounts** Not checked yet

NEW 85

Warms up agy CLI (Antigravity) account quota windows automatically on pool reset, with activity logged to `agy-warmup.json`, mirroring the existing ChatGPT warm-up behaviour. Enable it via the `agyWarmup` preference and track accounts through the new `agyWarmList` preference list in the provider editor's Advanced section.

— Names config file and both preference keys plus exact UI location v0.1.4671

03 **Per-tab visibility controls in Library** Not checked yet NEW 80

Hides individual Library tabs — Instructions, MCP servers, Skills, Hooks, Environment variables, Subagents, Slash commands, Rules, or Claude plugins — via the new `hide` setting in the Library plugin's Built-in settings (Plugins › Built-in), while agents still receive what those hidden tabs hold.

— Names exact config key, UI path, and every affected tab v0.1.4670

04 **Vibe Desktop detection without CLI install** Not checked yet NEW 60

Adds detection of Vibe Desktop (`Vibe.app` / `Vibe.exe` , app ID `ai.mistral.lechat.desktop`) to the `desktopApps` registry, so the Vibe agent is listed even when only the Mistral desktop app is installed and no CLI is present. A notice tells Vibe Desktop users that settings are read when a chat starts, so they must start a new chat (or restart open sessions) to apply a model switch.

— Names exact app files and ID but detection is passive, little user action v0.1.4676

05 **Tailscale MagicDNS remote connections** Not checked yet NEW

60

Lets a user connect to a remote magpie instance over plain HTTP by entering a Tailscale `.ts.net` hostname, without needing an HTTPS key.

— Clear mechanism but no further config or limits given v0.1.4672

THINNER COVERAGE BELOW

06 **Claude turn continuation after cancellation** Not checked yet

IMPROVED

45

Claude turns following a cancelled turn now continue in the same run, past Claude Code's 'No response requested.' local reply, instead of stopping.

— Behavior fix described but it's automatic, nothing to act on v0.1.4672

WAS THIS USEFUL?  

Alibaba Qwen Code

Feedback ⓘ desktop-v0.28.4 NOTES ↗

RANK

Qwen Code is a qwen offers comprehensive functionality spanning chatbot, image and video understanding, image generation, document processing, web search integration, tool utilization, and artifacts.

> Qwen Code expanded its managed messaging channels with new GitHub, WeCom and Microsoft Teams adapters plus generic webhook delivery, added several Web Shell UI surfaces (artifact library, Continue badge, suggestion chips, shared ChatSurface), and shipped a run of reliability fixes across Managed Agent telemetry, Autofix evidence display, Local Runtime permissions and extension manifest loading.

└──▷ **WHAT SHIPPED · 11 FEATURES** most completely described first ? what's the number?

01 **Channel adapters and webhook delivery** Not checked yet NEW

75

Adds `GitHub`, `WeCom`, and Microsoft Teams adapters to the managed channel path, joining `DingTalk` and other existing channels; the `WeCom` adapter handles inbound and outbound messaging and the Microsoft Teams adapter includes an extension and host ingress. Also adds generic outbound webhook delivery so any HTTP endpoint can receive channel output, with destinations configurable from the web UI.

— Names each adapter and mechanism plus webhook config surface. desktop-v0.28.4

THINNER COVERAGE BELOW

02 **Managed Agent session telemetry and object naming** Not checked yet

IMPROVED

55

Managed Agent sessions now emit a telemetry client-channel value so they can be distinguished from ordinary daemon traffic, and Managed Agent OSS object keys now use readable tenant, workspace, and session identifiers, making bucket contents easier to inspect.

— Names both telemetry field and object-key identifiers but no usage step. v0.28.4

03 **Continue badge for interrupted sessions** Not checked yet NEW

50

Web Shell adds a Continue badge in the sidebar for sessions whose last turn was interrupted and needs continuation.

— Names UI element and exact sidebar location. v0.28.4

04 **Shared ChatSurface chat column** Not checked yet NEW 50

Web Shell introduces a shared `ChatSurface` chat column that renders transcript, approvals, questions, and composer from session state.

— Names component and what it renders, with screenshots of thought display. v0.28.4

05 **Hosted Harness idle suspension** Not checked yet NEW 45

Implements stateless Hosted Harness phase 1 idle suspension, keeping idle MCP connections alive across suspension cycles.

— Names mechanism and scope but no user-facing action. desktop-v0.28.4

06 **Tappable suggestion chips for touch users** Not checked yet NEW 40

Web Shell composers gain tappable suggestion chips so touch users can insert and edit follow-up suggestions before sending.

— Describes behaviour but not exact composer location. v0.28.4

07 **Autofix evidence log sanitization** Not checked yet IMPROVED 40

Autofix failed-check evidence now strips terminal control characters and Unicode Bidi overrides from log excerpts before display.

— Specific sanitization targets named, no further action needed by user. v0.28.4

08 **Local Runtime state directory permission timing** Not checked yet IMPROVED 40

Local Runtime now creates its state directory at process spawn with owner-private permissions instead of during provider creation.

— States exact before/after timing and permission change. v0.28.4

09 **Fallback chain model availability tracking** Not checked yet IMPROVED 30

Tracks model availability for the fallback chain so the agent can reroute when a model is down.

— Explains purpose but no config or trigger detail. desktop-v0.28.4

10 **UTF-8 BOM extension manifest loading** Not checked yet IMPROVED

30

Extensions now load manifests saved with a UTF-8 BOM.

— Small named fix, minimal elaboration. v0.28.4

11 **External Shell worker profile and publication producer**

Not checked yet NEW

20

Adds an external Shell worker profile and a publication producer.

— Named but undescribed; no mechanism given. desktop-v0.28.4

WAS THIS USEFUL? 👍 👎

The Open Engine Zeroshot

Feedback ⓘ v17.1.1 NOTES ↗

RANK

Zeroshot orchestrates executor-verifier agent loops that implement and validate software changes while logging workflow state to SQLite.

> Zeroshot now supports self-hosting its target on Kubernetes via a versioned Helm chart.

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

01 Helm chart for self-hosted Zeroshot target

Not checked yet NEW

85

Deploy a self-hosted Zeroshot target on Kubernetes using the Helm chart at `oci://ghcr.io/the-open-engine/charts/zeroshot-target`. Each chart release pins the target image from the same release, and the `--version` flag selects an earlier release to install.

Install the Zeroshot target on a Kubernetes cluster, pinning to a specific earlier chart version rather than the latest stable.

```
$ helm install zeroshot-target oci://ghcr.io/the-open-engine/charts/zeroshot-target --version <version>
```

— Names exact chart OCI path, flag, and install command v17.1.1

WAS THIS USEFUL? 👍 👎

gentle-ai

Feedback ⓘ v5.0.0 NOTES ↗ RANK

> gentle-ai v5.0.0 adds a dedicated route for abandoning stuck review lineages and ships several breaking changes tied to its v4-to-v5 upgrade, including removal of legacy review authority and a silent capabilities payload change.

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

01 Review abandon command for stuck lineages

Not checked yet NEW

90

`gentle-ai review abandon` lets maintainers unstick a blocked review lineage through a single route, requiring `--expected-revision` to pass the exact current revision and `--maintainer-authorization` to supply the maintainer authorization binding.

Unstick a review lineage that is blocked and cannot progress normally, supplying the required revision and authorization credentials.

```
$ gentle-ai review abandon --expected-revision <revision> --maintainer-authorization <binding>
```

— Exact command and flags given, directly runnable. v5.0.0

02 Module and command paths moved to v5

Not checked yet BREAKING

65

The Go module and command paths move from `/v4` to `/v5`; integrators must update imports, build scripts, and pinned `go install` commands.

— Clear migration instruction, no deeper mechanism given. v5.0.0

THINNER COVERAGE BELOW

03 Capabilities v2.6 payload changed silently

Not checked yet

55

BREAKING

The `capabilities/v2.6` payload changed without an identifier bump, so consumers should refresh capability discovery rather than treating that identifier as proof of the old feature set.

— Names exact identifier but not what changed in payload. v5.0.0

04 Legacy v1 review authority removed

Not checked yet BREAKING

35

Legacy v1 review authority is removed; integrations that relied on it must switch to the supported review protocol.

— No detail on what the supported protocol requires. v5.0.0

└──▷ **BREAKING ON UPGRADE**

! The Go module and command paths move from `/v4` to `/v5`; update imports, build scripts, and pinned `go install` commands.

! Legacy v1 review authority is removed; integrations that relied on it must use the supported review protocol.

! The `capabilities/v2.6` payload changed without an identifier bump; consumers should refresh capability discovery rather than treating that identifier as proof of the old feature set.

WAS THIS USEFUL? 👍 👎

OpenQodex

Feedback ⓘ openqodex@0.22.1 NOTES ↗ RANK

OpenQodex is a join the world's most widely adopted, AI-powered developer platform where millions of developers, businesses, and the largest open source community build software that advances humanity.

> OpenQodex improved its `graph references` command with broader cast-pattern detection and smarter reporting when caller lists bottom out in test-only code.

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

01 **Test-only floor detection for caller lists** Not checked yet

IMPROVED 85

When a caller list is a floor and every candidate location is a test, fixture, corpus, benchmark, or top-level script, `unknown.scope` is now set to `tests` and `unknown.note` names up to three of those files; the CLI prints this as the sentence following `floor:`.

— Names exact fields and output behavior, explains mechanism openqodex@0.22.1

WAS THIS USEFUL? 👍 👎

Alibaba OpenCodeReview

Feedback ⓘ v1.14.1 NOTES ↗

RANK

> OpenCodeReview enriched its review findings with behavior and impact context, applying to both native and delegated review modes.

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

01 **Behavior and impact context in review findings** Not checked yet

IMPROVED 28

Review findings in both native and delegated modes now include behavior and impact information alongside each reported issue.

— Describes the change but no mechanism, UI path or example given. v1.14.1

WAS THIS USEFUL? 👍 👎

All Hands AI OpenHands

Feedback ⓘ

announced · Nov 12, 2026 BLOG ↗ RANK

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

> OpenHands launched Model Router inside Agent Canvas, letting conversations automatically pick the best LLM per task by cost and performance instead of pinning to one model.

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

01 **Model Router for LLM profiles in Agent Canvas** Not checked yet NEW

88

OpenHands adds a Model Router LLM profile type in `Settings > LLM` that routes each task to a different model — a strong reasoning model for risky refactors, a cheap one for routine work, a multimodal one for screenshots — without pinning a conversation to one model. Enable automatic per-conversation routing via the Run on first message toggle, which activates when a router is added, or ask the agent to invoke the router per step for intra-turn routing. Two ready-made routers ship — Router Pro (near-top accuracy at reduced cost) and Router Flash (40%+ cost reduction while outscoring several frontier single models) — plus a blank Custom template, each composed of an editable prompt template, a model table, and a classifier model (`MiniMax-M3`), all viewable and modifiable in `Settings > LLM`. LLM profiles can be created in bulk from a single provider API key via the router setup flow, and any running conversation can be switched to a specific LLM profile with the `/model` command.



Set up Router Flash for cost-optimized routing so every new conversation automatically picks the cheapest appropriate model.

📍 In Agent Canvas, go to Settings › LLM, select `Add Model Router`, choose `Router Flash`, connect your provider API key (the router creates all missing LLM profiles in bulk), then save. The `Run on first message` toggle activates automatically.

— Names exact UI path, toggle, command, and router types; mechanism detail is strong.

`launch-20261112-6cbbd469`

WAS THIS USEFUL? 👍 👎

◆ AI App Builders

Replit Agent

Feedback ⓘ announced · Nov 20, 2026 BLOG ↗

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

Replit Agent adds Tunnels, letting IT bridge agent workspaces to private-network services over mTLS without opening public inbound access.

NEW INTEGRATION Replit Agent adds Tunnels, a new capability that lets IT connect Replit environments to private-network services (databases, internal APIs) without opening public inbound access or exposing the whole network. `Not checked yet`

Create a tunnel in Workspace Settings and add endpoints so that all traffic — connector auth, Agent tool calls, and published-app requests — flows through Replit's tunnel server and a client you run inside your private network. `Not checked yet`

Tunnel traffic is encrypted and authenticated end-to-end using mTLS between the Tunnels client and Replit, with each private service's authentication and TLS configured separately by IT. `Not checked yet`

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

Arcade

Feedback ⓘ changelog · Nov 22, 2026 CHANGELOG ↗ RANK

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

> Arcade launched Multi-Profile support, letting a single gateway connect and route calls across multiple accounts per provider.

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

01 **Multi-Profile support across provider accounts** Not checked yet

NEW 79

A single gateway can now connect multiple accounts from the same provider (Google, Microsoft, Microsoft Power BI, or Slack), with agents selecting the target account per call via the new optional `arcade_profile_id` parameter on tool calls. Multi-Profile is enabled from the Organization section of the dashboard, is off by default, and leaves existing gateways unaffected until an org admin turns it on; every tool result now includes a note naming the account used, and all call records (Multi-Profile included) stream to the SIEM with the agent, user, and system identified under the same authorization and governance rules as before.

Activate Multi-Profile for your organisation so a single gateway can serve all connected provider accounts.

> 📍 In the Arcade dashboard, go to Organization settings and enable Multi-Profile.

— Names the parameter, providers, and dashboard toggle with clear setup steps.

launch-20261119-fd4f5380

NEW INTEGRATION Datadog toolkit

UPDATED INTEGRATION Linear toolkit

UPDATED INTEGRATION Gmail toolkit

RETIRED INTEGRATION Figma toolkit tools removed

changelog · Nov 22, 2026 CHANGELOG ↗ LATEST STABLE

> Arcade 1.13.0 adds connected-account profiles, org suspension controls, scoped/deactivatable API keys, a per-gateway Registry project catalog, and searchable admin lists.

Enable connected-account profiles so end-users can keep multiple connected accounts per integration and choose which one a tool call uses (opt-in per organization). Not checked yet

Suspend or unsuspend an organization, revoking its deployed workers, API keys, account tokens, and OAuth sessions in one command. Not checked yet

Deactivate, reactivate, and scope project API keys to a role (tools-only, manage-only, or full access). Not checked yet

Browse every toolkit available in a project through a per-gateway Registry project catalog toggle. Not checked yet

Search, filter, and sort Secrets, OAuth Provider, and Gateway admin lists with accurate totals. Not checked yet

+8 more

└──▷ BREAKING ON UPGRADE

- ! The `openaiApiKey` and `openaiApiKeySecretRef` chart values are deprecated and will be removed in chart 1.15.
- ! The `workers[].http` chart value is deprecated and will be removed in chart 1.15.
- ! The `supportImages.busybox` chart value has been removed (deprecated since chart 1.12).
- ! Scripts sending undocumented filter/sort values to list APIs now receive bad-request errors instead of being silently accepted.
- ! The unused `account` gateway binding type has been removed.
- ! `Draft` is no longer accepted as an MCP protocol version.
- ! Coordinator refuses to boot if its pinned authorization model drifts from the one shipped in its image.
- ! OAuth negotiation with servers that never confirmed PKCE support is now refused.
- ! Suspending an organization now immediately revokes its API keys, account tokens, and OAuth sessions.
- ! The Figma toolkit tools that caused Figma to reject sign-in for every Figma tool have been removed.
- ! The Telegram toolkit's `SendTtsAudio` tool is broken until the next toolkit update.

WAS THIS USEFUL? 👍 👎

open-multi-agent

Feedback ⓘ v1.25.0 NOTES ↗ RANK

open-multi-agent is a self-hosted TypeScript agent runtime with durable approvals and verifiable run records.

> open-multi-agent v1.25.0 adds a built-in Venice AI provider and fixes provider parameter forwarding, alongside several breaking changes to error handling in Gemini, content-filter finish reasons, and semaphore validation.

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

01 **Built-in Venice AI provider** Not checked yet NEW 90

Adds a built-in `venice` provider that defaults to `https://api.venice.ai/api/v1`, reads credentials from `VENICE_API_KEY`, and accepts a `baseURL` override in `AgentConfig`. Exhaustive switches on the `SupportedProvider` union must now add a `venice` branch or they will be incomplete.

Override the Venice endpoint (e.g. a corporate proxy) while still using `VENICE_API_KEY` for auth.

```json
{
  "provider": "venice",
  "baseURL": "https://proxy.example.com/venice/v1"
}
```

— Exact default URL, env var, config override, and runnable example given. v1.25.0

02 **Parameter forwarding fixes for Gemini and OpenAI/Azure**

Not checked yet IMPROVED 60

OpenAI and Azure OpenAI now send `max_completion_tokens` for GPT-5 and o-series models while other models continue to use `max_tokens`. Gemini's chat() and stream() now forward `topP` from `AgentConfig` instead of silently dropping it.

— Names exact params and models affected, no usage example. v1.25.0

THINNER COVERAGE BELOW

03 **Semaphore construction and release() validation** Not checked yet

55

BREAKING

Semaphore constructors now throw `RangeError` for `NaN` or below-1 limits, and an unmatched release() call now throws instead of silently corrupting the active count.

— Names exact error type and conditions, no code sample. v1.25.0

04 **Unified content_filter finish reason across providers**

Not checked yet

50

BREAKING

Consumers must now handle the `content_filter` finish reason for Anthropic refusals, Gemini prompt blocks, and OpenAI refusals, including OpenAI refusal text blocks.

— Lists affected providers but gives no handling example. v1.25.0

05 **Gemini now honors configured baseURL** Not checked yet

BREAKING

35

Gemini now dials the configured `baseURL` rather than ignoring it; teams should verify egress allowlists and proxy or gateway settings before upgrading.

— Names the config key but gives only a general warning. v1.25.0

└──▷ BREAKING ON UPGRADE

- ! Exhaustive switches on the `SupportedProvider` union must add a `venice` branch or they will be incomplete.
- ! Gemini now dials the configured `baseURL`; verify egress allowlists and proxy or gateway settings before upgrading.
- ! Consumers must now handle the `content_filter` finish reason for Anthropic refusals, Gemini prompt blocks, and OpenAI refusals (including OpenAI refusal text blocks).
- ! Semaphore constructors now throw `RangeError` for `NaN` or below-1 limits, and an unmatched release() call now throws instead of silently corrupting the active count.

WAS THIS USEFUL? 👍 👎

◆ Agent Sandboxes

sandboxd

Feedback ⓘ v0.3.28 NOTES ↗ RANK

`sandboxd` is a macOS daemon that enforces sandbox security policies by monitoring and denying unauthorized system calls made by sandboxed processes.

> sandboxd's v0.3.28 release introduces a configurable default AI model for chats, file attachment support across the console, and a unified agent entry point for all console access.

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

01 **Default AI model for new chats** Not checked yet NEW 70

Set a default AI model for all new chats using the `agents.default_ai` settings key, so every new chat session opens with the chosen model without manual selection each time.

Pin a specific AI model so every new chat session opens with it, without having to select it each time.

```yaml
agents.default_ai: gpt-4o
```

— Names exact config key and example value, but little mechanism detail. v0.3.28

THINNER COVERAGE BELOW

02 **File attachments in chat** Not checked yet NEW 25

Attach files on the Home screen and inside every app chat.

— No named surface, format, or limit given. v0.3.28

03 **Unified agent entry point in console** Not checked yet NEW 20

Access all console entry points through a single sandboxd agent front door.

— Vague description with no concrete mechanism or path named. v0.3.28

WAS THIS USEFUL? 👍 👎

◆ Browser & Computer Use

browser-use

Feedback ⓘ announced · Nov 20, 2026 BLOG ↗ RANK

The browser-use Python library automates web browsers for AI agents.

> browser-use 1.0 overhauls its agent architecture around a single persistent Python REPL tool with raw Chrome DevTools Protocol access, replacing the 24 discrete browser actions of 0.x, alongside a simplified three-line Agent API defaulting to Browser Use Cloud.

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

01 Python REPL tool with direct CDP access

Not checked yet BREAKING

87

The new `python` tool gives agents a persistent Python REPL with browser-harness helpers and cdp() pre-imported, letting code call any Chrome DevTools Protocol command directly via cdp('Domain.method', ...) for capabilities no built-in action covered. This replaces the 24 discrete browser action tools from 0.x entirely; agents relying on any of those named actions must be rewritten to use the `python` REPL tool with cdp() calls or custom `Tool` registrations.

Use raw CDP from inside the agent to unblock a hung page by stopping its load and blocking the stuck script before retrying.

python

```
cdp('Network.enable')
cdp('Network.setBlockedURLs', urls=['*<stuck-script>*'])
goto_url('https://<exchange-stats-page>/')
```

— Names the mechanism, lists what's removed, includes a runnable CDP example

launch-20261120-a4f9f6d6

02 Three-line Agent entry point defaulting to cloud

Not checked yet

NEW

66

Agent('...').run_sync(), imported via `from browser_use import Agent`, is now a three-line way to kick off an agent task, with the model and browser defaulting to Browser Use Cloud.

Kick off an agent task with the new three-line API, using the default Browser Use Cloud browser and model.

```python
from browser_use import Agent
result = Agent('Find the number of stars of the browser-use repo').run_sync()
print(result.final_result())
```

— Runnable example given, but cloud defaults not detailed further

launch-20261120-a4f9f6d6

THINNER COVERAGE BELOW

03 **Custom tool registration alongside python tool** Not checked yet

NEW

48

Additional custom tools can be registered alongside the built-in `python` tool using Tool(python, max_retries=TOOL_RETRIES, sequential=True).

— Names the syntax but no worked example or further mechanism launch-20261120-a4f9f6d6

└──▷ **BREAKING ON UPGRADE**

! The 24 discrete browser action tools from 0.x are removed; agents relying on any of those named actions must be rewritten to use the `python` REPL tool with cdp() calls or custom `Tool` registrations.

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Inference Providers & GPU Clouds

Anyscale

Feedback ⓘ changelog · Nov 22, 2026 CHANGELOG ↗ RANK

Anyscale provides a managed Ray platform for distributed AI development, training, batch processing, and model serving.

> Anyscale shipped cause-specific CLI exit codes, fixed workspace tag updates, and added support for custom headers on every CLI call.

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

01 Custom headers via ANYSCALE_EXTRA_HEADERS

Not checked yet NEW

73

Setting the `ANYSCALE_EXTRA_HEADERS` environment variable to a JSON object (e.g. `{"X-Proxy-Token": "abc"}`) passes those headers on every Anyscale CLI call, such as `anyscale job list`, without exposing the value in shell history.

Pass a proxy authentication header to every Anyscale CLI call without exposing it in your shell history.

```
$ export ANYSCALE_EXTRA_HEADERS='{"X-Proxy-Token": "abc"}'
anyscale job list
```

– Includes runnable example with exact env var and command changelog-20261122-3f0eb906

02 Workspace tags applied on update

Not checked yet IMPROVED

72

Workspace tags can now be applied when updating a workspace via anyscale.workspace.update() or `anyscale workspace_v2 update --config-file`; previously, tags specified in the config file were silently ignored.

– Names exact SDK call and CLI flag with before/after behavior

changelog-20261122-3f0eb906

03 Cause-specific CLI exit codes

Not checked yet IMPROVED

68

Failing commands now exit with cause-specific codes: `4` for permission (403) errors, `2` for unreachable API or server errors, `3` for builds exceeding one hour, and `1` for other failures, letting scripts distinguish failure causes programmatically.

– Lists exact codes and causes but truncated final category changelog-20261122-3f0eb906

WAS THIS USEFUL? 👍 👎

Together AI

Feedback ⓘ announced · Nov 21, 2026 BLOG ↗ RANK

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

> Together AI shipped a new text-to-speech model with streaming and WebSocket APIs, added Tencent's 770B-parameter Hy4 vision-and-tool-calling model with a 1M-token context, and introduced a free NVMe caching layer that speeds up storage reads on GPU Clusters by up to 20×.

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

01 Qwen3 TTS speech synthesis model

Not checked yet NEW INTEGRATION

NEW

95

Together AI now hosts Qwen3 TTS (`Qwen/Qwen3-TTS-12Hz-1.7B`) on Together Serverless Inference and dedicated endpoints at $20 per 1M characters. Generate speech via client.audio.speech.create(), controlling tone and emotion through the `instructions` field (e.g. excited and upbeat) passed via `extra_body` in the Python SDK. Nine preset voices each support Chinese, English, Japanese, Korean, German, French, Russian, Portuguese, Spanish and Italian, and audio can be returned in `wav` , `mp3` , `raw` (PCM) or `mulaw` formats at a `sample_rate` of 24000 (default) or 8000; `opus` , `aac` and `flac` are not supported.

Synthesize a line with a specific emotional delivery — here excited and upbeat — and save the result as an MP3.

python

```python
from together import Together
client = Together()
response = client.audio.speech.create(
    model="Qwen/Qwen3-TTS-12Hz-1.7B",
    input="Well, that changes everything.",
    voice="aiden",
    language="en",
    extra_body={"instructions": "excited and upbeat"},
    response_format="mp3",
)
response.write_to_file("excited.mp3")
```

— Full API call, model name, pricing, formats and voice/language list given

launch-20261120-c2bc072a

02 HTTP and WebSocket speech streaming

Not checked yet NEW

95

Stream synthesized audio over HTTP by setting `stream=True` and `response_format='raw'`; each `conversation.item.audio_output.delta` server-sent event carries a base64-encoded chunk of 16-bit mono PCM at 24 kHz. For real-time use, connect to `wss://api.together.ai/v1/audio/speech/websocket` and send text in pieces using `input_text_buffer.append` and `input_text_buffer.commit` messages, receiving audio as each sentence is synthesized.

Stream audio in real time over WebSocket when input text arrives token-by-token from a language model, committing once complete.

python

```
import asyncio, base64, json, os, websockets
URL = (
    "wss://api.together.ai/v1/audio/speech/websocket"
    "?model=Qwen/Qwen3-TTS-12Hz-1.7B&voice=aiden&language=en"
    "&response_format=pcm&sample_rate=24000"
)
async def main():
    headers = {"Authorization": f"Bearer
{os.environ['TOGETHER_API_KEY']}"}
    async with websockets.connect(URL,
additional_headers=headers) as ws:
        await ws.send(json.dumps({"type":
"input_text_buffer.append", "text": "Your order has shipped.
"}))
        await ws.send(json.dumps({"type":
"input_text_buffer.append", "text": "It should arrive on
Thursday."}))
        await ws.send(json.dumps({"type":
"input_text_buffer.commit"}))
        with open("order-update.pcm", "wb") as audio:
            async for message in ws:
                event = json.loads(message)
                if event["type"] ==
"conversation.item.audio_output.delta":

audio.write(base64.b64decode(event["delta"]))
                elif event["type"] ==
"conversation.item.audio_output.done":
                    break
asyncio.run(main())
```

— Exact endpoint, event names and message types with runnable example

launch-20261120-c2bc072a

03 Tencent Hy4 model with vision and tool calling

Not checked yet

NEW

85

Access Tencent Hy4 on Together Serverless Inference using the model string `tencent/Hy4` via the chat completions API, with a 1M-token context window at $0.83 per 1M input tokens, $0.042 per 1M cached input tokens, and $2.50 per 1M output tokens. It is the first model in the Hy family to accept images via an `image_url` block in message content alongside text, and supports structured output via `response_format: {"type": "json_object"}` or tool calling via the standard `tools` parameter.

Extract structured data from a chart image and return it as JSON using Hy4's vision and JSON output capabilities.

python

```python
import base64
from together import Together

client = Together()

with open("chart.png", "rb") as f:
    image = base64.b64encode(f.read()).decode()

response = client.chat.completions.create(
    model="tencent/Hy4",
    messages=[
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": "Read this chart. Return a JSON
object with a title and a list of bars, each with a label and a
value.",
                },
                {
                    "type": "image_url",
                    "image_url": {"url":
f"{image}"},
                },
            ],
        }
    ],
    response_format={"type": "json_object"},
)

print(response.choices[0].message.content)
```

— Model name, pricing and parameters given with working example `launch-20261119-923aeea4`

04 **Together Storage Accelerator for GPU Clusters** Not checked yet NEW

85

Together Storage Accelerator (TSA) is a distributed caching layer that runs on every GPU node in a Together GPU Cluster, serving warm reads from local NVMe at up to 3.4 GB/s — roughly 20× faster than cross-cloud origin pulls — and fetching each object from origin only once regardless of how many nodes request it. Point S3-compatible storage (S3, R2, T4) through it by setting `AWS_ENDPOINT_URL=http://tsa.tsa.svc.cluster.local:18080`, or route Hugging Face Hub downloads by setting `HF_ENDPOINT=http://tsa.tsa.svc.cluster.local:18080`, with no code changes to existing boto3 or CLI workflows; it is enabled from a cluster's Settings page at no additional cost.

Route all S3-compatible reads (S3, R2, T4) through TSA's local cache so that the first node to fetch an object populates the NVMe cache for every other node in the cluster.

```
$ export AWS_ENDPOINT_URL=http://tsa.tsa.svc.cluster.local:18080
aws s3 cp s3://my-bucket/checkpoint.bin ./checkpoint.bin
```

— Exact env vars, speed figures and enablement path given launch-20261119-a45f770f

05 **VoiceFinder voice preview tool** Not checked yet NEW

60

Preview and compare all nine Qwen3 TTS preset voices against custom text (up to 500 characters) using VoiceFinder at `voicefinder.together.ai`, with generated code snippets in Python, TypeScript or cURL for the chosen voice.

Audition all nine Qwen3 TTS voices against your own text before committing to one for your application.

📍 1. Open `voicefinder.together.ai`. 2. Select `Qwen/Qwen3-TTS-12Hz-1.7B` in the model menu, then select Find voices. 3. Select Preview next to any voice to hear a sample. 4. Select Try custom text, enter up to 500 characters of your own text and select Generate. 5. Repeat for other voices to compare on the same text. 6. Select Show code to copy a ready-made Python, TypeScript or cURL request for the chosen voice.

— Exact URL and character limit but UI-only workflow launch-20261120-c2bc072a

announced · Nov 21, 2026 BLOG ↗

Together AI adds Qwen4 Max with a 1M-token context window, streaming-only inference, built-in thinking controls, image input, tool calling, and automatic prompt caching.

Access Qwen4 Max via Together Serverless Inference using the model string `Qwen/Qwen4-Max`, which requires `stream=True` (or `"stream": true` in the request body) — requests without streaming return a 400 error with code `streaming_required`. Not checked yet

Control thinking on Qwen4 Max by passing `reasoning={"enabled": False}` to disable it, or setting `reasoning_effort` to `"max"` for harder problems; thinking is on by default and streams in the `delta.reasoning` field, with thinking tokens counted as output tokens. Not checked yet

Qwen4 Max accepts text and image input, with images sent as `image_url` blocks in the message content. Not checked yet

Qwen4 Max supports tool calling through the standard `tools` parameter and can return more than one call per turn. Not checked yet

Prompt caching is automatic for Qwen4 Max on Together AI, with cached input tokens priced at `$0.075` per 1M tokens at list price (or `$0.0525` during the sale through November 22, 2026). Not checked yet

---

WAS THIS USEFUL? 👍 👎

◆ ML Platforms & SDKs

AWS Amazon Bedrock AgentCore

Feedback ⓘ

changelog · Nov 22, 2026 | CHANGELOG ↗ | RANK

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

> Amazon Bedrock AgentCore shipped per-event streaming interception and custom domains for Gateways, skill packages in the AWS Agent Registry, an expanded Runtime Instances compute type with nested KVM virtualization and G7 GPU instances, S3 bucket mounts for agent sessions, and new dataset generation and Insights capabilities in AgentCore Evaluations.

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

01 S3 bucket mounts for agent sessions

Not checked yet | NEW

90

Agent sessions can mount a shared S3 bucket via `fileSystemConfig.s3BucketMount`, specifying `bucketName`, `prefix`, and `mountPath` — each session can write into its own prefix folder so runs do not collide.

Mount a shared S3 bucket into every agent session, with each session writing into its own prefix folder so runs do not collide.

json

```
{
  "fileSystemConfig": {
    "s3BucketMount": {
      "bucketName": "my-agent-data",
      "prefix": "sessions/{runtimeSessionId}/",
      "mountPath": "/mnt/data"
    }
  }
}
```

— Full config example with per-session prefixing changelog-20261122-d024ccc9

02 Expanded Instances compute type for AgentCore Runtime

Not checked yet | IMPROVED

75

The Instances compute type of AgentCore Runtime gains elevated-privilege agent execution, nested KVM virtualization for running virtual machines in agent sessions (supporting Android emulators and sandboxed code execution), and support for G7 EC2

instances (`g7.2xlarge` through `g7.48xlarge` , with NVIDIA RTX PRO 4500 Blackwell GPUs) added to a capacity provider's allowed instance types.

– Names instance types and capabilities but no config key shown

changelog-20261122-d024ccc9

03 **Custom domain support for JWT-authorized gateways** Not checked yet

NEW

60

Gateways behind a proxy can now use a custom domain via the new `domainOverride` setting, with the `x-amzn-bedrock-agentcore-gateway-domain` header controlling which configured domain applies per request.

– Names config field and header but no example changelog-20261122-d024ccc9

WAS THIS USEFUL? 👍 👎

◆ Inference Serving & Optimization

Niko1221 Strata

Feedback ⓘ v0.2.19 NOTES ↗ RANK

Strata is a local inference engine with one-click install for Windows and Linux, providing an OpenAI/Anthropic-compatible API on localhost for models like Qwen and DeepSeek.

> Strata's v0.2.19 release adds native Inkling 2 Mini support with a dedicated shape engine, ships three quantization packs and a vision adapter for it, bundles speculative decoding, and speeds up Qwen3.8-Flash-Next inference on several GPU families.

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

01 **Native Inkling 2 Mini support via strata-inkling engine**

Not checked yet NEW 90

Inkling 2 Mini can be loaded locally using `strata launch` with `--gpus 2` for Dual-GPU Expert Parallelism, or loaded directly via `strata-inkling` from official `safetensors` or community `.gguf` files. This works across Windows, Linux, macOS, and NVIDIA DGX Spark.

– Names exact launch command and flag, plus supported file formats. v0.2.19

02 **Strata Quantization packs for Inkling 2 Mini** Not checked yet NEW 80

Three Strata Quantization packs are available at `strata-engine/Inkling-2-Mini-Strata` on Hugging Face for cards with 4 GiB or more: SQ-10 (10.47 GB), SQ-15 (14.35 GB, 96.68% top-1 agreement), and SQ-20 (19.35 GB, lossless SF4+MX8, 100% top-1 agreement).

– Exact repo path and per-pack sizes/accuracy given, but no load command. v0.2.19

03 **Vision adapter and DSpark speculative decoding for Inkling 2 Mini**

Not checked yet NEW 65

A lossless BF16 vision adapter, `Inkling-2-Mini-Vision.strata` (0.64 GB), ships alongside a built-in `W=4` Multi-Branch On-GPU Markov Beam Tree DSpark speculative decoding feature, both working out of the box in the desktop app and via `strata launch`.

– Names the adapter file and decoding mechanism but no setup steps. v0.2.19

04 **15 ready-made strata-inkling shape engine packages** Not checked yet

NEW 65

Strata now provides 15 ready-made `strata-inkling` 0.1.65 packages in `shape_engines`, fetched automatically on demand, covering CUDA 12/13, HIP RDNA2/RDNA3/RX9070/Halo, Metal (macOS arm64), SYCL (Intel Arc), and Linux aarch64 (DGX Spark GB10).

— Lists all target platforms and version but fetch is automatic, no user action. v0.2.19

THINNER COVERAGE BELOW

05 **Faster Qwen3.8-Flash-Next inference** Not checked yet IMPROVED

55

Strata now delivers faster Qwen3.8-Flash-Next inference on NVIDIA Blackwell, RTX 40, RTX 30, and AMD RDNA3 with zero environment variables needed, with up to +76% prefill speed and 2.91x faster TTFT on RTX 5090.

— Gives concrete speedup numbers but improvement is automatic, no action needed. v0.2.19

WAS THIS USEFUL? 👍 👎

OpenNMT CTranslate2 Feedback ⓘ v4.9.1 NOTES ↗

CTranslate2 runs Transformer models with a C++ inference engine, quantization, and CPU and GPU optimizations.

CTranslate2 v4.9.1 makes `flash_attention` effective for the Whisper encoder.

└──▷ GET THIS VERSION ▾

pip

```
$ pip install ctranslate2==4.9.1
# already installed? upgrade to this version:
$ pip install --upgrade ctranslate2==4.9.1
```

git

```
$ git clone --branch v4.9.1
https://github.com/OpenNMT/CTranslate2.git
# already have the repo? check out this version:
$ git checkout v4.9.1
```

named in the project's README · checked 22 Nov Wrong command?

NEW SUPPORTED PLATFORM NVIDIA Blackwell GPU targets in CUDA 12 builds

WAS THIS USEFUL? 👍 👎

vLLM

Feedback ⓘ announced · Nov 20, 2026 BLOG ↗ RANK

The vLLM inference engine runs large language models with high-throughput, memory-efficient GPU execution.

> vLLM ships day-0 support for Inkling 2 Mini with speculative decoding, adds architecture-level support for looped-layer models with configurable pass control, and integrates with SkyRL for reinforcement-learning serving.

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

01 Looped-layer model support with variable pass control

Not checked yet NEW 91

vLLM now supports looped-layer models where middle layers run multiple passes with shared weights, per-pass KV and conv-state caches, and gated message reads between passes. For the TML-d24-varloop-research checkpoint, the number of loop passes and KV-cache sharing can be set at startup via `--hf-overrides` using the `loop_num_passes` and `loop_share_cache` keys inside `text_config`.

Run the variable-loop research checkpoint with 2 passes and a shared KV cache to trade some quality for faster decode and larger KV cache capacity.

```
$ docker run --gpus all --ipc=host -p 8000:8000 \
  -v ~/.cache/huggingface:/root/.cache/huggingface \
  vllm/vllm-openai:inkling-2-mini thinkingmachines/TML-d24-varloop-research \
  --reasoning-parser inkling --tool-call-parser inkling --enable-auto-tool-choice \
  --hf-overrides '{"text_config": {"loop_num_passes": 2, "loop_share_cache": true}}'
```

— Names config keys and gives a runnable startup command launch-20261120-1af1e96b

02 Day-0 Inkling 2 Mini serving support

Not checked yet NEW 90

vLLM adds day-0 support for Inkling 2 Mini, served via the `vllm/vllm-openai:inkling-2-mini` Docker image with `--reasoning-parser inkling`, `--tool-call-parser inkling`, `--enable-auto-tool-choice`, and DSpark speculative decoding configured via `--speculative-config '{"method":"dspark","num_speculative_tokens":7,"draft_sample_method":"probabilistic"}'`.

Serve Inkling 2 Mini on a Blackwell or Hopper GPU with reasoning, tool-calling, and 7-token DSpark speculative decoding enabled.

```
$ docker run --gpus all --ipc=host -p 8000:8000 \
  -v ~/.cache/huggingface:/root/.cache/huggingface \
  vllm/vllm-openai:inkling-2-mini thinkingmachines/Inkling-2-Mini \
  --reasoning-parser inkling --tool-call-parser inkling --enable-auto-tool-choice \
  --speculative-config '{"method":"dspark","num_speculative_tokens":7,"draft_sample_method":"probabilistic"}'
```

— Runnable command with exact flags and image name given `launch-20261120-1af1e96b`

THINNER COVERAGE BELOW

03 **Reinforcement-learning serving via SkyRL** Not checked yet NEW

25

vLLM adds reinforcement-learning serving of Inkling 2 Mini via SkyRL integration.

— Only named, no mechanism or usage details given `launch-20261120-1af1e96b`

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

Pinecone

Feedback ⓘ announced · Nov 19, 2026 BLOG ↗ RANK

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

> Pinecone added a Cost and Usage Export API that delivers billing data in FOCUS v1.4 CSV format for direct ingestion into FinOps tooling.

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

01 **Cost and Usage Export API** Not checked yet NEW 88

Export cost and usage data via the `cost-usage-exports` endpoint of the Admin API, returning a zip of FOCUS v1.4 CSV files tagged by project and index, ready for FinOps tools without manual reshaping. Exports can cover any window in the past year, be created from the Usage page in the console, or scheduled for automated pulls into a data warehouse or FinOps pipeline, eliminating manual spreadsheet downloads.

Create a cost and usage export through the console to see Pinecone spend broken down by project and index in existing FinOps reports.

📍 1. In the Pinecone console, go to Usage. 2. Select the date range (up to one year in the past). 3. Click Create export to download a zip of FOCUS v1.4 CSV files.

— Names exact endpoint, format, and console steps launch-20261119-218320f9

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI Security Posture & Discovery

Mend.io Mend AI

Feedback ⓘ announced · Nov 19, 2026 BLOG ↗

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

Mend AppSec gains Red Hat Lightwell integration: backported-fix detection, advisory reconciliation, and Mend Renovate presets for Java and Python.

NEW SUPPORTED PLATFORM Configure Mend Renovate to ingest Red Hat Lightwell backported builds for Java and Python using the published presets in the Red Hat Lightwell configuration repository. Not checked yet

NEW INTEGRATION Mend AppSec now detects installed Red Hat Lightwell builds and automatically reconciles findings against Red Hat Lightwell advisories, keeping transitive and residual risks visible. Not checked yet

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Anthropic

Feedback ⓘ changelog · Nov 22, 2026 CHANGELOG ↗ RANK

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

> Anthropic's Claude Developer Platform shipped customer-linked billing, automatic prompt caching on Bedrock, new encryption-key webhook events, the Opus 4.7 model with breaking tokenizer changes, text watermarking, cross-platform access via an OpenAI-compatible endpoint and Amazon Bedrock, and expanded Skills, web fetch, and vision capabilities.

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

01 Automatic prompt caching on Bedrock and Prompt Caching API

Not checked yet NEW 95

Add `bedrock-mantle` to your request body to enable automatic prompt caching on Amazon Bedrock, which moves the cache point forward as conversations grow without manual breakpoint management. The Prompt Caching API is available elsewhere to cache and reuse prompt content, reducing latency by up to 80% and costs by up to 90%.

Automatically cache the last cacheable block on Amazon Bedrock as a conversation grows, avoiding manual cache-point management.

```
$ curl https://<bedrock-endpoint>/v1/messages \
  -H 'content-type: application/json' \
  -d '{
    "model": "claude-opus-4-6",
    "max_tokens": 1024,
    "bedrock-mantle": true,
    "messages": [{"role": "user", "content": "Continue the
analysis..."}]
  }'
```

— Named field, mechanism, and quantified latency/cost benefits with example

changelog-20261122-8b0fa230

02 Customer-linked billing via billing_policy

Not checked yet NEW 90

Control who pays for API usage by setting `billing_policy` on each `/v1/messages` or `/v1/chat/completions` request after a customer links their organization, using the `user-profiles-billing-2026-10-01` beta header.

Let a customer's Claude organization pay for API usage instead of yours — useful for SaaS apps where each tenant has their own Anthropic contract.

```
$ curl https://api.anthropic.com/v1/messages \
  -H 'x-api-key: $ANTHROPIC_API_KEY' \
  -H 'anthropic-version: 2023-06-01' \
  -H 'anthropic-beta: user-profiles-billing-2026-10-01' \
  -H 'content-type: application/json' \
  -d '{
    "model": "claude-opus-4-6-20261101",
    "max_tokens": 1024,
    "billing_policy": "customer",
    "messages": [{"role": "user", "content": "Summarize this
report."}]
  }'
```

— Named endpoints, field and beta header plus runnable curl example

changelog-20261122-8b0fa230

---

03 **OpenAI-compatible endpoint and Amazon Bedrock availability**

Not checked yet NEW 75

The Claude API now has an OpenAI-compatible endpoint, letting you test Claude models by changing only your API key, base URL, and model name in existing OpenAI integrations. Claude is also now available on Amazon Bedrock at a new endpoint that uses the same request shape as the first-party Claude API and runs on AWS-managed infrastructure with zero operator access.

— Clear migration path described but no concrete endpoint or command given

changelog-20261122-8b0fa230

---

04 **Claude Opus 4.7 model release** Not checked yet BREAKING 70

Claude Opus 4.7 is now available at $5 / $25 USD per MTok, with an updated tokenizer and API breaking changes versus Opus 4.6; consult the Migration guides before upgrading.

— Pricing and breaking change noted but no migration steps given

changelog-20261122-8b0fa230

---

05 **Encryption key webhook events** Not checked yet NEW 60

New `external_key.failed` and `external_key.recovered` webhook event types (in beta) let you track when a customer-managed encryption key is refused or recovers.

— Named event types but no usage example given changelog-20261122-8b0fa230

THINNER COVERAGE BELOW

06 **File format support in Prebuilt Skills** Not checked yet IMPROVED 50

Prebuilt Skills now support PowerPoint (`.pptx`), Excel (`.xlsx`), Word (`.docx`), and PDF files.

– Names exact formats but no usage mechanism or example changelog-20261122-8b0fa230

07 **Text watermarking for generated content** Not checked yet NEW 40

Text generated by Claude Opus 4.6, Opus 4.5, Sonnet 4.5, and Haiku 4.5 now carries Anthropic's text watermark on supported platforms.

– States which models gained it but no mechanism or opt-in detail

changelog-20261122-8b0fa230

08 **Web fetch tool (beta)** Not checked yet NEW 40

The web fetch tool (in beta) lets Claude retrieve full content from specified webpages and PDF documents.

– Describes function but lacks invocation details or example changelog-20261122-8b0fa230

09 **Vision support on Claude Haiku 3.5** Not checked yet NEW 35

Vision support is now available on Claude Haiku 3.5, enabling the model to analyze images.

– Bare capability statement with no example or mechanism changelog-20261122-8b0fa230

10 **Memory for Claude Managed Agents** Not checked yet NEW 25

Memory for Claude Managed Agents is now available in beta.

– Bare announcement with no mechanism or usage detail changelog-20261122-8b0fa230

└──▷ **BREAKING ON UPGRADE**

! Claude Opus 4.7 includes an updated tokenizer and API breaking changes versus Opus 4.6; see the Migration guides before upgrading.

WAS THIS USEFUL? 👍 👎

◆ Open-Weight Models

Ai2 Models

Feedback ⓘ announced · Nov 19, 2026 BLOG ↗ RANK

Ai2 develops the Olmo, Molmo, Tulu, and OlmoEarth families of open language and multimodal models.

> Ai2 introduced Theorizer, a new AI research agent in its Asta platform that generates evidence-linked scientific theories, alongside an open-source release for self-hosted use.

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

01 **Theorizer agent in Asta** Not checked yet NEW 55

Theorizer is a new AI agent available in Asta that searches the research literature, extracts structured findings, and proposes evidence-linked explanations for a user's scientific question. Ai2 has open-sourced the Theorizer codebase so researchers can examine how it generates theories and adapt the methods, including a self-hosted path to run the library on their own machines and pay for tokens directly.

— Describes mechanism and self-hosting but no concrete command or API given.

launch-20261119-7dd36a9c

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · November 22, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue