THE AI TOOLCHAIN

NO. 057

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED OCTOBER 11, 2026 · EVERY DAY

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

▸ // HOW THIS ISSUE IS MADE SHOW

VIEW

ISSUE VIEW

scroll



SCROLL | EXPAND IN PLACE | PREVIEW PANE

INFORMATION DENSITY

everything



BRIEFED | EVERYTHING

Every tool in the issue, in full.

all 136 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 21 tools

▾ // THE BRIEF HIDE

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

Three unrelated things stand out. The OpenAI Python SDK can now authenticate with a Kubernetes service-account token or mutual-TLS so no API key sits in your cluster. LiteLLM makes OpenTelemetry

vim ? help

prompt/response capture a per-tenant decision rather than all-or-nothing. And Open WebUI adds enforceable MFA with a CLI recovery path for self-hosted deployments.

GOVERN

Call OpenAI from a pod without storing an API key anywhere

Workload identity auth lets the SDK present a Kubernetes service-account token or an X.509 client cert instead of a long-lived secret, so there is no key to rotate, leak in an env var, or sync into every namespace. HTTPX2 also becomes the default HTTP client.

OpenAI Python SDK

GOVERN

Log full prompts for one tenant and only metadata for another

LiteLLM makes OpenTelemetry prompt/response capture a per-tenant setting, which is what you need when one customer's contract forbids payload retention but you still want traces for everyone else. Open WebUI adds enforceable MFA with a CLI recovery path, so an admin lockout does not require a database edit.

LiteLLM · Open WebUI

BUILD

Keep a long-running coding agent session alive across failures

Qwen Code adds managed session failover and session resume in the TypeScript SDK, plus session-centric multi-agent collaboration — a dropped connection no longer discards the agent's accumulated context. OpenCode adds Agent Skill metadata controls in the TUI for picking which skills a run can reach.

Qwen Code · OpenCode

OPERATE

See what a Mistral-backed app actually costs without hand-maintaining a price table

Phoenix adds Mistral AI pricing to cost tracking, so spans from those models carry a dollar figure instead of raw token counts you convert yourself. magpie adds per-key gateway usage limits, capping a runaway agent at the key rather than at the invoice.

Arize Phoenix

BUILD

◆ AI Coding Agents

yetone magpie

Feedback ⓘ v0.1.1195 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.

> Across 21 releases, magpie widened from a model switcher into a broader agent-and-routing hub — expanding supported agents to 50+ (adding StepCode, Prime Agent, jcode, and others as full integrations), adding provider surfaces (You.com search, Tencent Cloud Token Plan 企业版, ModelScope via Responses API), and shipping new routing primitives (folder-based rules, a decision-model group at `/v1/systemone`, routing-group text sharing) alongside per-key gateway usage limits, MCP server management, and more reliable Windows detection for Claude Code and Desktop.

└──▷ **WHAT SHIPPED · 59 FEATURES** most completely described first ? what's the number?

01 Per-key usage limits for remote magpie gateways NEW

90

Adds a `GET /v1/magpie/limit` endpoint (`RemoteKeyLimitPath`) on a remote magpie gateway so each gateway key can fetch its own limit, used, remaining, period, and reset time — never another key's numbers. The Usage page now shows per-key usage cards for remote magpie providers, separate from the remote's subscription cards.

Query a remote magpie gateway to check your own key's spending limit and remaining budget — authenticated with that key alone.

```
$ curl -H 'Authorization: Bearer <your-gateway-key>'
https://<remote-magpie-host>/v1/magpie/limit
```

— Named endpoint, method, path, with a runnable curl example v0.1.1179

02 MCP server management: per-tool toggles, removal, listing

NEW

85

Adds per-tool enable/disable switches for each MCP server's tools in the library and GUI, writing the disabled list into each agent's own config format — Codex `disabled_tools`, Gemini CLI `excludeTools`, Crush `disabled_tools`, Claude Code `permissions.deny` (`mcp__<server>__<tool>`), and OpenCode's `tools` map in `opencode.json`. Enables removal of an agent-owned MCP server from that

agent, with automatic backup of the original config file before modification, and lists an MCP server's tools by name and description in the library and GUI views.

— Names every agent's own config key for disabled tools `v0.1.1183` `v0.1.1191` `v0.1.1179`

03 **Plugin API library wiring and MCP formats** NEW 85

Adds a `library` field to agent plugin exports, with `instructions`, `mcp`, `mcpFormat`, and `skills` sub-fields, so a plugin-added agent can receive the Library's instructions file, MCP servers, and skills folder at the paths it names. Supports MCP file formats `alma`, `antigravity`, `claude`, `cline`, `codex`, `commandcode`, `copilot`, `crush`, `cursor`, `devin`, `gemini`, `goose`, `grok`, `hermes`, `kimi`, `omp`, `opencode`, and `zed` via `agent.library.mcpFormat`. Also exposes the agent's working folder and git-repository status to a plugin's fetch call, without exposing magpie's own folder.

Wire a plugin-added agent (e.g. Aider) so magpie automatically populates its instructions file, MCP server list, and skills folder from the Library.

shell

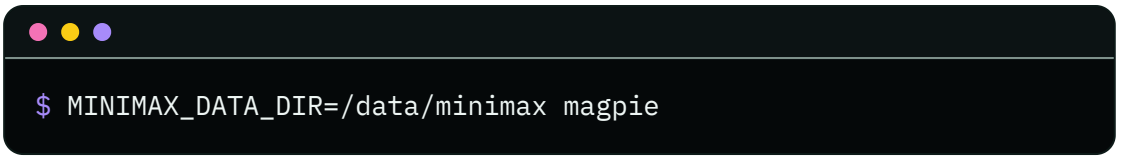
```
export const agent = {
  id: "aider",
  name: "Aider",
  dir: "~/.aider",
  config: "~/.aider/config.yml",
  model: "gpt-4o",
  library: {
    instructions: "~/.aider/AGENTS.md",
    mcp: "~/.aider/mcp.json",
    mcpFormat: "claude",
    skills: "~/.aider/skills",
  },
};
```

— Names field, sub-fields, formats, with a runnable manifest example `v0.1.1177` `v0.1.1191`

04 **MiniMax Code usage tracking** NEW 85

Tracks MiniMax Code's own API calls in Usage's per-agent split by reading `@minimax-ai/code` 0.6.8 session files; reads the data dir from the `MINIMAX_DATA_DIR` env var, defaulting to `~/.minimax`.

Point magpie at a non-default MiniMax Code data directory so Usage's per-agent split correctly attributes MiniMax Code calls.

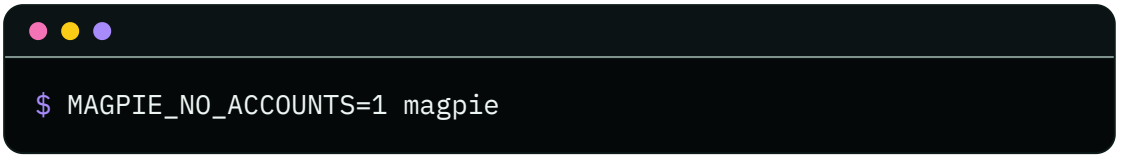


— Named env var, default path, package version, runnable example v0.1.1188

05 **Disable subscription account discovery** NEW 85

Subscription account discovery can now be disabled so that no agent sign-in credentials are read for any agent. Set `MAGPIE_NO_ACCOUNTS=1` to prevent magpie from reading Claude Code or Codex sign-in files without changing the saved setting, or run `magpie accounts discovery off` to permanently disable discovery so magpie never touches agent sign-in files or the Keychain.

Prevent magpie from reading Claude Code or Codex sign-in files to discover subscription accounts, without changing the saved setting — useful in CI or shared environments.



Permanently disable subscription account discovery so magpie never touches agent sign-in files or the Keychain.

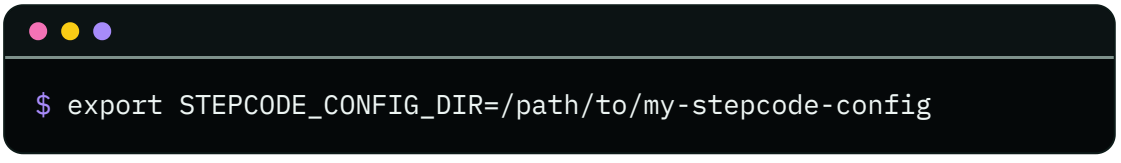


— Names env var and exact command, both runnable v0.1.1176

06 **StepCode agent integration** NEW 85

Adds StepCode agent integration: `stepcode` can be connected to magpie, with configuration managed in `~/.stepcode/config.toml` (or the directory set by `STEP_CODING_AGENT_DIR` / `STEPCODE_CONFIG_DIR`), provider entry written to `models.json`, and Disconnect restoring the original model.

Point StepCode at a custom config directory so magpie writes its provider entry into the right `models.json`.



— Names config paths, env vars, and a runnable example v0.1.1172

07 **Decision-model routing group at /v1/systemone** NEW 85

Adds a decision-model routing group that answers at `/v1/systemone`, failing over between models in routing order — addressable as `group/<id>` or by group id or name.

Send a decision-model request to a named routing group, automatically failing over through the group's members.

```
$ curl -X POST http://127.0.0.1:3425/v1/systemone \
  -H 'Content-Type: application/json' \
  -d '{"model":"group/<id>","messages":
[{"role":"user","content":"Classify this alert."}]}'
```

— Named endpoint, addressing scheme, and runnable example `v0.1.1170`

---

08 **Search API providers, routing, and MCP exposure** NEW

80

Adds You.com as a search API provider, supporting both keyless free search and authenticated access with an API key. Every search API now accepts a per-key custom base URL, configurable under its key in Settings. Magpie's web search is exposed as an MCP server at `SearchMCPPath` (port 3425), consumable by any MCP-compatible client, and web searches are now routed through the configured searcher's own proxy rather than the provider's connection, isolating search traffic from LLM provider traffic.

Add You.com free web search as a provider without needing an API key, giving agents live search results at no cost.

> 📍 In Settings › Providers, click Add Provider and select You.com; leave the API key blank to use its free search tier.

— Names provider, config location, MCP port, and routing behavior `v0.1.1170` `v0.1.1184` `v0.1.1191`

---

09 **jcode agent integration** NEW

80

Adds `jcode` as a connectable agent: magpie registers itself as the `providers.magpie` profile in jcode's `config.toml` (located in `$JCODE_HOME` or `~/.jcode`), exposing all catalog models in jcode's model picker and starting jcode on the selected model.

— Names config path and env var, no CLI command shown `v0.1.1173`

---

10 **Session ID interpolation in provider Extra body** NEW

75

Adds `${sessionId}` interpolation to a provider's Extra body field, injecting JSON fields into every request body with the current session ID.

Inject a stable per-session identifier into every outgoing request body — useful for tracing or provider-side session correlation.

📍 In Settings › Providers, open a provider and set Extra body to: {"x-session": `${sessionId}` }

— Named field with a working configuration example v0.1.1191

11 **--help flag and model-value validation hardening** IMPROVED

75

Passing `--help` or `-h` to any command or agent now prints only that command's usage and exits, rather than treating the flag as a value — previously `magpie snow --help` set Snow's model to `--help` and corrupted the active profile. Snow and Gemini CLI agents now refuse unknown model values outright instead of silently accepting them and destabilising magpie state.

Inspect the usage for the snow agent without accidentally mutating its model setting.

```
$ magpie snow --help
```

— Before/after bug description with a runnable command v0.1.1185

12 **Tencent Cloud Token Plan 企业版 support** NEW

75

Adds Tencent Cloud Token Plan 企业版 as a provider region (Enterprise plan · China and · Global), authenticated with the account's `SecretId` and `SecretKey`. Surfaces its usage on the Usage page, read via `DescribeTokenPlanList` and `DescribeTokenPlan` on `tokenhub.tencentcloudapi.com` (API version `2026-03-22`, signed `TC3-HMAC-SHA256`).

— Names auth fields, API methods, host, and version v0.1.1182

13 **/v1/codex/models endpoint for Fast Codex models** NEW

75

Exposes Fast model offers from the Codex CLI via `/v1/codex/models` on a provider's GPT model.

Query the Codex models endpoint to see which Fast-tier GPT models magpie is offering through the Codex CLI provider.

```
$ curl http://localhost:<gateway-port>/v1/codex/models
```

— Named endpoint with a runnable curl example v0.1.1172

14 **Git-commit-aware plugin update detection** IMPROVED 70

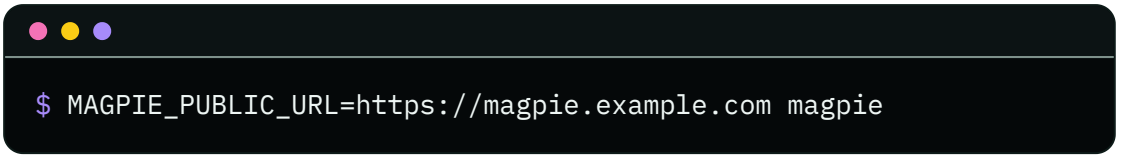
Plugins installed from a git repository now show `Update` only when the repository has new commits, comparing the commit recorded in `node_modules/<pkg>/.bun-tag` (or `bun.lock`) against the current branch/tag head via GitHub's API or `git ls-remote`.

— Names exact files and comparison methods used v0.1.1169

15 **MAGPIE_PUBLIC_URL for gateway address** NEW 65

Adds `MAGPIE_PUBLIC_URL` environment variable so Gateway's Connect starts at that URL by default and `magpie web` displays it as the gateway address.

Expose a stable public gateway URL so remote users connecting via the web UI always see and start from a known address.



— Named env var with a runnable example v0.1.1181

16 **Prime Agent integration** NEW 65

Adds Prime Agent as a first-class agent row on the Agents page, wired like Pi in `~/.prime/agent` (with `settings.json` and `models.json`); `Disconnect` restores both files byte-for-byte.

— Names exact config paths, no CLI example given v0.1.1172

17 **Folder-based routing rules** NEW 60

Routing-group rules can now match requests by the working directory the agent operates in, with support for Claude Code, Codex, Pi, and OpenCode agents. Pause rules can likewise be scoped by folder — held inside specific folders, or held outside them.

— Names supported agents, no exact config steps shown v0.1.1180 v0.1.1191

18 **Grok context-window variant models** NEW 60

Exposes each Grok context-window variant (e.g. `grok-4.7@500k`) as a selectable model alongside the base model, matching what the CLI's `/context-window` command lists.

— Names example model id and matching CLI command v0.1.1188

19 **Claude Code /model lists magpie routing groups** NEW

60

Claude Code's `/model` command now lists magpie's routing groups, so a saved routing group is selectable as the active model.

— Names exact command and clear usage path v0.1.1182

20 **Routing-group sharing as text** NEW

60

Adds routing-group sharing as text using a `ShareMark` / `ShareVersion` format, so a group's rules, members, and routing can be copied from one magpie instance and pasted into another.

— Names format, no exact command or UI path shown v0.1.1181

21 **Gateway conversation recording for Pi and omp** IMPROVED

60

Extends Record gateway conversations to cover Pi and `omp` agent turns, alongside `OpenCode`, and sessions recorded by the gateway now download as Markdown instead of returning a 400.

— Names agents and fixes download format, no exact path v0.1.1169

THINNER COVERAGE BELOW

22 **Codex subagents route to any provider's model** IMPROVED

55

Expands `CodexSubagentModel` so Codex subagents can be assigned any provider's models (not only ChatGPT models), routing `collab_spawn` requests across the full set magpie serves.

— Names config field and request type, no usage steps given v0.1.1195

23 **Minis compaction prompt support in routing groups** NEW

55

A routing group's compaction rule now accepts Minis's compaction prompts, matching OpenMinis Android `ChatViewModel.compactSummarySystemPrompt` and iOS `AIChatViewModel+Compaction.swift`.

— Names exact source files, no usage steps given v0.1.1194

24 **Gateway handling of relay busy responses** IMPROVED

55

Gateway now correctly treats a relay's 500 busy server response as a transient load condition and waits it out using the group's Patience setting, rather than counting it as a quota error.

— Names Patience setting and before/after behavior v0.1.1188

25 **ModelScope routed via Responses API for Claude Desktop** IMPROVED 55

Routes Claude Desktop requests to ModelScope through ModelScope's Responses API instead of its Chat API, surfacing cache and reasoning fields previously omitted.

— Names API switch and surfaced fields, no exact steps v0.1.1183

26 **Gzip compression and caching for magpie web** IMPROVED 55

Serves gzip-compressed pages and JSON from `magpie web`, with ETag-based 304 caching for versioned scripts — reducing roughly 4 MB of scripts and styles transferred over slow links.

— Names mechanism and savings figure, no config needed v0.1.1171

27 **Claude account quota check without switching** IMPROVED 50

Refreshing a saved Claude account's card now reads its allowance through Claude Code without switching to that account, avoiding spent refresh tokens.

— Clear mechanism, no exact command or UI path v0.1.1194

28 **Long-task completion notifications** NEW 50

Adds opt-in long-task completion notifications via Settings › Usage, alerting when an agent's turn finishes.

Get notified when a long-running agent task completes so you can switch windows and return only when work is done.

📍 In Settings › Usage, enable the long-task notification opt-in.

— Exact settings path given, limited further mechanism v0.1.1191

29 **Stalled image response diagnostics** IMPROVED 50

A stalled image response now reports whether it was arriving as a `url` or `b64_json`, how much had been received, and how long it had been stalled.

— Names response fields, no UI path given v0.1.1188

30 **Expanded agent integrations list** NEW 50

Extends agent support to include Gemini CLI, Antigravity CLI, OpenCode, OpenChamber, MiMo Code, Pi, oh-my-pi, Aside, OmO, Prime Agent, Goose, Cursor CLI, Cursor Private Inference, Zed, VS Code Chat, VS Code Insiders, VSCodium Chat, JetBrains Air, Copilot (JetBrains), Copilot CLI, Crush, DeepSeek Harness, Reasonix Studio, Command Code, fx, Devin, Hermes Agent, Mister Morph, Kimi Code, jcode, Qwen Code, Muse Code,

Empryo, MiniMax Code, Droid, Cline, Qoder, Qoder CN, Grok Build, ZCode, WorkBuddy, WorkBuddy AI, CodeBuddy Code, Pencil, T3 Code, OpenHanako, AtomCode, Snow CLI, Alma, and Cindy.

— Lists every supported agent, no mechanism or steps `product docs`

---

31 **Per-group speed labels with reply count** IMPROVED 50

Shows per-group speed labels in the UI that include the reply count — e.g. `125 tok/s (1)` for a group with replies, or `≥100 tok/s (30)` for one with floor-only data.

— Gives concrete example values, no deeper mechanism `v0.1.1180`

---

32 **Live API and model validation in GUI** NEW 50

Adds live API and model validation in the GUI: picking a model or clicking `Detect APIs` checks the form as it currently stands (before any Save) and surfaces missing fields inline.

— Names UI action, no deeper mechanism given `v0.1.1173`

---

33 **Copilot CLI detection on WSL and Linux keyring** IMPROVED 50

Supports finding the Copilot CLI signed in inside a running WSL distro on Windows, and discovers the Copilot CLI 1.0 token from the Linux keyring.

— Names token source and environment, no steps given `v0.1.1172`

---

34 **Search boxes added to Providers and Settings lists** NEW 45

Adds a search box to the Providers list in the GUI, with a right-click context menu to move a provider to the top or bottom, and a separate search box to the Settings panel that filters settings by name or partial keyword match.

— UI locations named, no deeper mechanism described `v0.1.1191` `v0.1.1186`

---

35 **Remote magpie instances grouped separately in Add sheet** IMPROVED 45

Remote magpie instances now appear under their own `Another computer's magpie` heading in the Add sheet, separated from Relays.

— Names exact UI heading, no further mechanism `v0.1.1185`

---

36 **Plugin-provided System One classifier models** NEW 45

A plugin can now serve System One for its provider, and models it marks become a routing group's classifier.

— Describes mechanism without a concrete example `v0.1.1169`

37 **Release publishing and update flow** IMPROVED 40

Each tagged release is now automatically copied to `yetone/magpie` with its notes and all 16 assets, and the GUI update flow opens the release page in the browser when an update is available, downloading, or failed.

— Internal process change, no user action required v0.1.1194

38 **Codex GPT model availability via relay** NEW 40

Codex GPT models are now offered via magpie when Codex is on a relay or when multiple accounts are configured in magpie.

— States condition, no deeper mechanism described v0.1.1188

39 **Marketplace search resilience** IMPROVED 40

Marketplace search now names which source failed to respond, shows results from the remaining sources, and automatically retries a 503 once.

— Describes behavior, no reproducible steps v0.1.1188

40 **dsh per-protocol API routing** IMPROVED 40

Extends `dsh` agent integration so each magpie model uses its own API route per protocol, routing requests across separate provider APIs rather than a single shared route.

— States mechanism, no configuration details given v0.1.1173

41 **Faster Requests tab over long usage histories** IMPROVED 40

Improves Requests tab, filters, charts, TUI, and `magpie usage` response time over long usage histories by loading all at once.

— Names affected surfaces, no numbers given v0.1.1173

42 **Bulk skill removal via 'All' chip** NEW 40

Adds an `All` chip on the skill-agent assignment screen to remove a skill from all agents in one click.

— Names exact UI control and location v0.1.1170

43 **Gemini-only custom provider model queries** IMPROVED 40

Allows a custom provider's model to be queried with Gemini alone, beside chat, Responses, and Anthropic.

— Names protocols, no exact config steps given v0.1.1169

44 **Inspect exact request text from context window** NEW

40

Clicking a cell or item in a context window opens the exact text sent in that request.

— Names exact UI interaction, no further detail v0.1.1169

45 **LAN sharing key delivery to WSL agents** IMPROVED 40

Turning LAN sharing on now delivers the sharing key to a WSL agent under NAT that was connected before sharing was enabled, resolving 401 errors.

— Names fixed error code, no reproduction steps v0.1.1169

46 **Automatic model selection limited to enabled models** IMPROVED 35

Restricts automatic model selection for image recognition and web search to only models the user has explicitly enabled, preventing routing to unenabled models.

— Behavior change described, no named surface or steps v0.1.1195

47 **Fast model switching for user-added provider GPT models** NEW 35

Enables fast model switching for GPT models on a user-added provider directly from Routing and the model picker.

— Brief description, no exact steps given v0.1.1191

48 **Direct LAN routing for plugin requests** IMPROVED 35

Routes plugin requests to hosts on the user's own network directly, bypassing magpie's proxy, matching magpie's own behaviour.

— States behavior, no configuration surface named v0.1.1184

49 **4K relay image fallback to URL** IMPROVED 35

Saves 4K images answered by URL from a relay, and falls back to returning the URL rather than losing the image when download fails.

— Describes fallback behavior, no named config v0.1.1184

50 **ZCode model picker filtering parity** IMPROVED 35

ZCode's agent row now filters its model picker to only the models selected for it on the Agents page, matching the behaviour of Pencil, T3 Code, WorkBuddy, and Copilot in JetBrains.

— Names agents matched, no deeper mechanism v0.1.1171

51 **Icon fetching proxied to hide user IP** IMPROVED 35

Fetches remote icons through `usemagpie.ai/api/icon` so the icon host never sees the user's IP address.

– Names proxy endpoint, no further mechanism v0.1.1170

52 **WorkBuddy credit-expiry-aware routing** IMPROVED 30

Smart routing now spends WorkBuddy credits that expire soonest first.

– Brief statement, no mechanism or steps given v0.1.1188

53 **Unresponsive model-list key kept active** IMPROVED 30

A key whose model-list endpoint did not respond is still kept active and accepts requests, rather than being excluded as unlisted.

– Thin behavior note, no mechanism detail v0.1.1188

54 **Magpie leads macOS Command-Tab order** IMPROVED 30

Ensures magpie leads Command-Tab application order on macOS when its window is open, so one Command-Tab away and back returns to it.

– Describes OS behavior, no configuration involved v0.1.1186

55 **In-place masking-rule editing** IMPROVED 30

Enables in-place editing of masking rules, rather than requiring removal and re-entry.

– Thin description, no named surface v0.1.1181

56 **Cached agents render immediately on startup** IMPROVED 30

Renders previously cached agents immediately on the Agents page after startup or a light-mode reopen, then replaces them with a fresh read.

– Describes behavior, no measurable detail v0.1.1170

57 **On-demand system proxy allowance reads** IMPROVED 30

Adds on-demand allowance reads so a system proxy set after magpie starts is picked up within seconds.

– Brief statement, no named surface v0.1.1169

58 **Plugin update button styling** IMPROVED 25

Plugin update button is now the same size as the Install button, styled in green, and appears on the plugin card, list row, and Update all view.

– Cosmetic change, no functional detail v0.1.1188

59 **Localized agent advice after config changes** IMPROVED 20

Agent advice after a configuration change now renders in the window's current language.

— Brief statement without further detail v0.1.1171

WAS THIS USEFUL? 👍 👎

SST OpenCode

Feedback ⓘ v1.19.0 NOTES ↗ CODE ↗ RANK

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

> OpenCode shipped new Agent Skill metadata controls for the TUI and overhauled its update flow, adding a one-step major-version upgrade path while requiring explicit manual updates going forward.

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

01 Major version upgrade and manual update command

BREAKING

81

`opencode upgrade --major` migrates from OpenCode 1 to OpenCode 2 and removes OpenCode 1 in one step, regardless of the package manager used to install. Automatic update installation has been replaced by the `/update` TUI command, so updates no longer install automatically and must be triggered manually from within the TUI.

Migrate an existing OpenCode 1 installation to OpenCode 2 in a single command, using whichever package manager you installed with.

```
$ opencode upgrade --major
```

— Runnable command plus clear breaking-change migration note. v1.19.0

02 Skill exposure metadata keys NEW

77

New `opencode/slash` metadata key (set to `"true"`) exposes an Agent Skill as a `/name` slash command in the TUI. New `opencode/autoinvoke` metadata key (set to `"false"`) excludes a skill from model-facing auto-selection guidance while still allowing manual selection by exact name.

Expose a skill as a slash command in the TUI so you can invoke it manually without relying on model auto-selection.

```yaml
metadata:
  opencode/slash: "true"
  opencode/autoinvoke: "false"
```

— Names exact config keys and values with a working example. product docs

└──▷ **BREAKING ON UPGRADE**

! Updates are no longer installed automatically; you must run `/update` inside the TUI to apply them.

WAS THIS USEFUL? 👍 👎

Token Optimizer

Feedback ⓘ v5.13.37 NOTES ↗ RANK

Token Optimizer is a coding-agent plugin that trims wasted context, checkpoints work across sessions and compactions, and reports token spend on a live dashboard.

> Token Optimizer's v5.13.37 release adds environment-aware status bar reporting, usage-based compaction advice on the dashboard, and tightens `kill-stale` session termination safety.

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

01 **Safer kill-stale session termination logic** IMPROVED

70

`kill-stale` now acts only when it can prove a session is not the one it is currently running in, and only operates on the current user's own sessions.

— Names the exact command and the new safety conditions v5.13.37

02 **Desktop status bar reads compact window env var** NEW

65

The desktop status bar now reads `CLAUDE_CODE_AUTO_COMPACT_WINDOW` and displays a single compact line by default, expandable for more detail with a link to the full dashboard.

— Names the env var and UI behavior but not exact config path v5.13.37

THINNER COVERAGE BELOW

03 **Usage-based compaction advice on dashboard** NEW 45

The dashboard now surfaces advice, based on your own session history, on whether a different compact window or a 1-hour subagent cache would reduce token costs.

— Describes the advice concept without a concrete UI path or config v5.13.37

WAS THIS USEFUL? 👍 👎

Diagram Design

Feedback ⓘ changes since 2026-09-28 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 delta diagrams, exploded axonometric, and axonometric plan diagram types.

└──▷ GET THIS VERSION

```
$ git clone --branch commits-2026-09-28
https://github.com/cathrynlavery/diagram-design.git
# already have the repo? check out this version:
$ git checkout commits-2026-09-28
```

└──▷ TRY IT

Rebuild all exploded axonometric examples from the projection model after editing coordinates or skins.

```
$ python3 scripts/build-exploded-examples.py
```

Verify that no exploded axonometric example file is stale relative to the current model, suitable for CI gating.

```
$ python3 scripts/build-exploded-examples.py --check
```

Adds architecture delta diagrams for visualizing structural changes between system states.

Adds the exploded axonometric diagram type for rendering separated component views in 2:1 dimetric projection.

Adds the axonometric plan diagram type for top-down axonometric layouts.

Prepares a Codex manifest for OpenAI's plugin directory, enabling the library to be listed and invoked as an OpenAI plugin.

Adds example diagrams for AI agent stack, keyboard, coffee shop, and warehouse in axonometric style.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ sdk-typescript-v0.1.19

NOTES ↗ RANK

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

> Qwen Code added an experimental Kubernetes CSI-backed runtime and session-centric multi-agent collaboration to its agent runtime, shipped a Batch API workflow and managed session failover for the desktop app, and extended the TypeScript SDK with multi-modal input, session resume, and new configuration keys, while the Web Shell picked up a large set of trajectory, host, and session management tools.

└──▷ **WHAT SHIPPED · 21 FEATURES** most completely described first ? what's the number?

01 Web Shell trajectory, host and session management tools

IMPROVED 80

Web Shell gains Qwen Live access with native macOS Host opt-in; host controls for model management; time-range selection on the trajectory overview to narrow the session table; in-conversation search; inspection of session tool calls by prompt; selective host artifact integration; Git worktree management from the git dialog; zoom and pan controls on the trajectory overview with display of trajectory over real time including idle periods; trajectory search and diagnostic filters for inspecting agent execution history; the ability to create independent side-task conversations in a parent session's secondary workspace, including while the parent is responding; a Send now action for queued mid-turn messages; support for browsing earlier trajectory windows; fullscreen viewing for history tables; connecting saved daemons simultaneously with no-flash host switching; and restoring aggregated workspace hosts in the sidebar.

— Many named UI capabilities, but only UI paths, no commands v0.25.1 desktop-v0.25.1

02 TypeScript SDK session and provider configuration NEW

78

Adds a `settings.env` field for environment variable configuration directly in settings; adds `resume` and `continue` options and extended `authType` support to the TypeScript SDK; adds support for resuming sessions with a session ID via the `query` interface; adds `extra_body` support for OpenAI-compatible providers; and adds a `contextWindowSize` configuration key.

Inject environment variables for a session without modifying your shell profile — useful for per-project API keys or proxy settings.

```json
{
  "env": {
    "OPENAI_API_KEY": "sk-...",
    "HTTPS_PROXY": "http://proxy.corp.example.com:8080"
  }
}
```

Resume a previous coding session by ID to continue exactly where you left off — useful after an interrupted run.

```
$ qwen --resume <session-id> -p "Continue the refactor from where we stopped"
```

— Names every config key and SDK option with runnable examples `sdk-typescript-v0.1.19`

---

03 **Multi-agent team and child-agent infrastructure** NEW

75

Adds the `H4a` child agent and child acceptance record contract, `H4b` child Session runtime, `H4c` workflow child kind and child launch budget, `H4d-a` session message record contract and child continuation rules, `H4d-b` session message runtime, `H4e-a` team record contract, `H4e-b1` lead-side team runtime, `H4f` public task cancel endpoint for child agent tasks, worktree admission for managed child agents, and child Workspace capability for managed child Sessions.

— Enumerates internal contracts; one public endpoint named `v0.25.1`

---

04 **CLI session management commands** IMPROVED

75

Improves CLI session management with the new `qwen sessions ps` command to list managed Agent View sessions, and adds support for model and reasoning selection at session creation time in `qwen serve`.

List all active managed Agent View sessions to audit what is currently running before attaching or cleaning up.

```
$ qwen sessions ps
```

— Exact command named with a runnable example `desktop-v0.25.1`

---

05 **Kubernetes CSI-backed runtime for managed sessions** NEW

70

Adds an experimental Kubernetes CSI-backed runtime that validates workspaces, persists pod identity, and records durable worker acknowledgements for managed sessions, and adds background Shell and Monitor capabilities to the Managed Agent runtime, enabling durable background shell sessions and task event reporting.

— Detailed mechanism and scope, but no CLI command or flag given `v0.25.1`

06 **Managed Runtime attestation and session failover** NEW

65

Adds a managed runtime attestation client to the Java SDK requiring attestation before a Managed Runtime is adopted, the `managed-context/1` envelope contract to the CLI for scoped context passing, and a Durable Managed Session journal and failover for resilient agent sessions.

— Names the contract and client but no usage steps `desktop-v0.25.1`

07 **Agent channel routing, access policies, and email delivery (H5 stage)**

NEW

65

Adds planned Stage `H5` agent channel API route shapes and persistence tables for channel routes and outbound deliveries, the `H5a` channel route and delivery record contract, and the `H5b` / `H5c` channel runtime with an email reference adapter for managed-agent channel delivery, and separates private and group channel access policies while routing channel messages to prefix-specific sessions.

— Names internal stage contracts but no public API shape No config keys or endpoints named

`v0.25.1` `desktop-v0.25.1`

08 **PreToolUse hook with input revalidation** NEW

65

Adds `PreToolUse` hook support with `updatedInput` and full revalidation, letting hooks modify tool inputs before execution.

— Named hook and field, no example of usage `v0.25.1`

09 **Workspace lifecycle management (W1c, L3, replay-floor)** NEW

60

Adds offline workspace migration (`W1c`), reliable ACTIVE Workspace deletion (`L3`), opt-in production replay-floor advancement for Session event retention, and a durable, idempotent way for creators to change a Workspace-bound Managed Session's working directory, including via WebShell, for managed agents.

— Named stage IDs, no exact command or endpoint given `v0.25.1`

10 **Batch execution workflow and concurrent runner** NEW

60

Adds a `/batch-api` agent-prepared Batch API workflow for running batch jobs from within a session, and a concurrent runner for batch CLI execution.

Kick off a Batch API job from within an interactive session to process large workloads without blocking the terminal.

```
$ /batch-api
```

— Named slash command with a runnable example No flags or invocation named

desktop-v0.25.1 sdk-typescript-v0.1.19

11 **disable* settings renamed to enable* naming** BREAKING

60

Negative `disable*` settings are renamed to positive `enable*` naming; any config keys using the old `disable*` form must be updated.

— Clear migration note but no exact key list given sdk-typescript-v0.1.19

12 **QWEN_DESKTOP_DISABLE_UPDATES environment variable** NEW

60

Adds the `QWEN_DESKTOP_DISABLE_UPDATES` environment variable to opt out of the startup update check.

Prevent the desktop app from checking for updates on every launch — useful in locked-down or air-gapped environments.

```
$ export QWEN_DESKTOP_DISABLE_UPDATES=1
```

— Exact env var name with a runnable example desktop-v0.25.1

THINNER COVERAGE BELOW

13 **Automation scheduling runtime (H6 stage)** NEW 50

Adds the H6a schedule and automation run record contract and the H6b / H6c automation runtime for persistent automation definitions and scheduling.

— Names stage contracts with no usage instructions v0.25.1

14 **Hosted Workspace /2 profile glob discovery** NEW 50

Adds read-only glob file discovery to Hosted Workspace `/2` profiles while keeping searches inside the session working directory.

— Named profile path, no command example v0.25.1

15 **Multi-modal input across content generators** NEW 45

Adds multi-modal input support (image, PDF, audio) across all content generators.

– Names modality types but no usage path `sdk-typescript-v0.1.19`

16 **Session-centric multi-agent collaboration** IMPROVED 40

Enables session-centric multi-agent collaboration, replacing the thread backend with A2A sessions.

– High-level architecture change with no migration detail `v0.25.1`

17 **Session history export command** NEW 40

Adds an export command for session history with markdown and HTML formats.

– Names output formats but not the command itself `sdk-typescript-v0.1.19`

18 **Screen sharing in voice calls** NEW 35

Streams shared screens into browser voice (Live) calls.

– No UI path or command given `desktop-v0.25.1`

19 **Zed extension for Qwen Code agent server** NEW 35

Adds a Zed extension for the Qwen Code agent server.

– Names the integration target but no setup steps `sdk-typescript-v0.1.19`

20 **Native advisor tool** NEW 30

Adds a native advisor tool with consultation behavior and usage limits.

– Vague on mechanism and usage limits `desktop-v0.25.1`

21 **Experimental LSP support** NEW 25

Adds experimental LSP support for code intelligence.

– No protocol scope or commands named `sdk-typescript-v0.1.19`

└──▷ **BREAKING ON UPGRADE**

! Negative `disable*` settings are renamed to positive `enable*` naming — any config keys using the old `disable*` form must be updated.

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

OpenClaw

Feedback ⓘ v2026.10.1 NOTES ↗ CODE ↗ RANK

OpenClaw is a personal AI assistant platform that connects messaging channels, tools, and model providers through a gateway.

> OpenClaw's v2026.10.1 release adds inline YouTube playback and workboard sidebar pinning to the Control UI, persists code-mode values across sessions, and optimizes session-list loading.

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

01 **Persistent code-mode store/load values** NEW 55

Code mode now keeps `store` / `load` values across cells, replies, and restarts, so state persists between sessions instead of resetting.

— Names exact commands and describes persistence scope v2026.10.1

02 **Optimized sessions.list loading** IMPROVED 55

The gateway now serves compact `sessions.list` rows and pushes changes to clients instead of requiring a full refetch, reducing session-load overhead.

— Names the endpoint and mechanism, but no metrics given v2026.10.1

03 **Workboards in the sidebar with pinning** NEW 40

Workboards now appear in the sidebar, and users can pin their own workboards for quick access.

— Clear UI location but no further detail on behavior v2026.10.1

04 **Inline YouTube playback in Control UI** NEW 35

OpenClaw now plays YouTube videos inline in Control UI chats, rather than requiring an external link.

— Names the surface but no mechanism or limits given v2026.10.1

05 **Web-style identity footer on macOS sidebar** NEW 25

The native macOS sidebar now shows a web-style identity footer.

— Bare description with no further mechanism or purpose v2026.10.1

└──▷ **ALSO FROM THESE RELEASES**

WAS THIS USEFUL? 👍 👎

assistant-ui

Feedback ⓘ @assistant-ui/react@0.15.27 NOTES ↗

RANK

Assistant-ui provides React components and state management for chat interfaces with streaming, tool calls, and attachments.

> Assistant-ui tightened thread-disabled state handling and expanded MCP App initialization data in a single patch release.

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

01 **MCP App initialize params passed to onInitialized** IMPROVED 62

An MCP App's `ui/initialize` params are now passed to `onInitialized` as `McpAppInitializeParams`, including the display modes the app declares.

— Names exact type and callback but no usage example @assistant-ui/react@0.15.27

THINNER COVERAGE BELOW

02 **Full input lockout on disabled threads** IMPROVED 30

Blocks typing, attachments, sends, and message editing while a thread is disabled.

— States behavior but no API or flag named @assistant-ui/react@0.15.27

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

HeyGen HyperFrames

Feedback ⓘ v0.8.146 NOTES ↗ RANK

HyperFrames is an open-source HTML-to-video renderer that runs in AI-agent workflows.

> HyperFrames v0.8.146 adds a new audio limiter effect and speeds up re-renders by reusing previously extracted video frames.

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

01 **Frame reuse on re-renders** IMPROVED 45

Re-renders now reuse the extracted frames of remote videos, eliminating redundant network fetches on repeated exports.

— Explains behaviour but no config or metric given v0.8.146

02 **Truepeak lookahead true-peak limiter FX** NEW 35

A new Truepeak lookahead true-peak limiter effect is available as an FX option in HyperFrames.

— Named but mechanism and usage unspecified v0.8.146

WAS THIS USEFUL? 👍 👎

emisar

Feedback ⓘ v0.53.0 NOTES ↗ RANK

Emisar lets AI agents securely run infrastructure actions through MCP, with approval controls that protect production systems.

> Emisar v0.53.0 deepens its infrastructure-read modules with approval-gated ClickHouse DDL access, IPv6-aware TLS inspection, and refined diagnostics across PostgreSQL, Traefik, Cassandra, Consul, Cloudflare, GCP Storage, and Linux, while retiring older module versions by default and shipping a standalone Linux amd64 diagnostics artifact.

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

01 **IPv6 TLS inspection and probe overrides in Network/TLS**

IMPROVED 70

Network/TLS 0.3.2 adds IPv6 inspection, bounded `hostname:port:IP` overrides for HTTP probes, headers that preserve Host, SNI, and certificate verification, legacy TLS probes, bounded WHOIS-to-RDAP fallback, and certificate inspection that now emits only the presented certificate chain. Versions below 0.3.1 are retired.

— Lists concrete probe override format and TLS behaviors, no runnable command shown.

v0.53.0

02 **ClickHouse structural and DDL reads** IMPROVED 65

ClickHouse 0.2.18 adds a normalized structural read that omits literals and is not recreatable SQL, plus a full server-provided DDL read that requires approval under the default policy and warns about sensitive defaults, comments, and engine arguments.

— Describes two read types and the approval gate, no command given. v0.53.0

THINNER COVERAGE BELOW

03 **Standalone Linux amd64 diagnostics artifact** NEW 55

A separate native Linux amd64 diagnostics artifact now carries its complete runtime closure and can be extracted by the host without starting a container.

— Names artifact and behavior but gives no size or checksum detail. v0.53.0

04 **PostgreSQL VACUUM and checkpoint diagnostics** IMPROVED 55

PostgreSQL 0.2.21 adds native VACUUM dead-item counters across versions, labeled checkpoint statistics, and a strict sequential-scan filter that includes indexless tables.

— Names specific counters and filters but no query or command example. v0.53.0

05 **Cassandra client metadata and storage read confinement**

IMPROVED

55

Cassandra 0.6.10 adds client metadata and confines storage reads to configured roots with keyspace-escape rejection and bounded compaction history. Versions below 0.6.10 are retired and blocked by default.

— Lists several mechanisms but no config surface named. v0.53.0

06 **Legacy integration versions retired and blocked by default**

50

BREAKING

Consul versions below 0.2.38, Cloudflare below 0.2.10, Cassandra below 0.6.10, Grafana below 0.1.20, and Network/TLS below 0.3.1 are retired and blocked by default unless explicitly overridden by an administrator.

— Names exact retired versions and the admin override path. v0.53.0

07 **Traefik metrics entrypoint and router inventories** IMPROVED

45

Traefik 0.1.25 adds a separate metrics entrypoint and complete bounded HTTP router/service inventories, including summary and readiness joins.

— Names entrypoint and inventory joins but no config path given. v0.53.0

08 **Consul token metadata masking** IMPROVED

40

Consul 0.2.38 omits SecretID values from token metadata, and Runner masking now recognizes the observed spellings. Versions below 0.2.38 are retired and blocked by default.

— Short note on masking, no mechanism detail beyond spelling match. v0.53.0

09 **Cloudflare Pages value redaction and ruleset checks** IMPROVED

40

Cloudflare 0.2.11 omits Pages environment values, preserves bounded API failures, and confirms an absent custom-ruleset entrypoint before reporting zero. Versions below 0.2.10 are retired and blocked by default.

— Lists three behaviors but no config key named. v0.53.0

10 **Linux peer counts and group protocol states** IMPROVED

35

Linux Core 0.5.3 and Debugging 0.3.4 add peer counts and group protocol states with inline port-set membership resolution.

— Names new data points but no scope or command detail. v0.53.0

11 **GCP Storage live vs noncurrent version distinction** IMPROVED

25

GCP Storage 0.1.3 distinguishes live and explicitly requested noncurrent object versions.

— Single-sentence capability with no mechanism detail. v0.53.0

└──▷ **BREAKING ON UPGRADE**

! Consul versions below 0.2.38, Cloudflare below 0.2.10, Cassandra below 0.6.10, and Grafana below 0.1.20 are retired and blocked by default unless explicitly overridden by an administrator.

! Network/TLS versions below 0.3.1 are retired.

WAS THIS USEFUL?  

OpenAI Python SDK

Feedback ⓘ v3.29.0 NOTES ↗ RANK

The OpenAI Python SDK provides typed Python clients for OpenAI model APIs, streaming, and asynchronous requests.

> The OpenAI Python SDK's biggest change this window is a shift to HTTPX2 as the default HTTP client alongside new zero-secret workload identity authentication (Kubernetes service-account tokens and mutual-TLS/X.509), plus promotion of the Responses API and a new Live voice.

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

01 **HTTPX2 is now the default HTTP client** BREAKING 57

The SDK switches its default HTTP client to HTTPX2. Any custom HTTP client, transport, timeout, authentication handler, event hook, or request mock configured against the old HTTPX client must be migrated — see the `httpx2.md` guide.

— Names the migration guide but no exact migration steps given. product docs

02 **New Live voice: shida** NEW 23

Adds `shida` to the named Live voices.

— Bare one-line addition with no usage detail. v3.29.0

└──▷ **ALSO FROM THESE RELEASES**

Use aiohttp as the async transport for higher concurrency throughput in an async pipeline.

python

```python
import asyncio
from openai import AsyncOpenAI, DefaultAioHttpClient

async def main():
    async with AsyncOpenAI(
        http_client=DefaultAioHttpClient(),
    ) as client:
        response = await client.responses.create(
            model="gpt-5.5",
            input="Analyse this log for anomalies.",
        )
        print(response.output_text)

asyncio.run(main())
```

└──▷ **BREAKING ON UPGRADE**

! The default HTTP client is now HTTPX2; any custom HTTP client, transport, timeout, authentication handler, event hook, or request mock configured against the old HTTPX client must be migrated — see `httpx2.md`.

WAS THIS USEFUL?  

◆ Local LLM Runtimes

Open WebUI

Feedback ⓘ v0.12.0 NOTES ↗ RANK

Open WebUI is a self-hosted chat interface for local and remote models, with RAG over uploaded documents, tool calling, and multi-user access control.

> Open WebUI's v0.12.0 release and accompanying docs add realtime voice/video calls with VRM avatars, enforceable MFA with a CLI recovery path, an experimental Tool Search mode for large toolsets, and a Community hub for sharing presets, skills, tools, and chat snapshots, alongside version history, nested groups, and a redesigned knowledge-base file browser.

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

01 **Realtime voice and video calls with avatars** NEW 100

Adds hands-free voice and video calls, realtime voice calls, and speech-to-text/text-to-speech support, configured via `AUDIO_REALTIME_ENABLED`, `AUDIO_REALTIME_OPENAI_API_BASE_URL`, `AUDIO_REALTIME_OPENAI_API_KEY`, `AUDIO_REALTIME_MODEL`, `AUDIO_REALTIME_VOICE`, `AUDIO_REALTIME_TRANSCRIPTION_MODEL`, and `REALTIME_CALL_PROMPT_TEMPLATE` environment variables. Small talk routes through an OpenAI Realtime voice model (`gpt-realtime-2.1-mini` with the `Marin` voice by default) while real questions go to the chat's selected model; switched in Admin Settings > Audio > Voice calls from Standard to Realtime, requires the `Allow Call` permission, and calls end after one hour. Also adds realtime animated VRM avatars to these calls — upload a rigged VRM file (up to 25 MiB) and optional VRMA animation clips (up to 10 MiB each) for idle, listening, and speaking states, plus up to 16 named gestures, in the model editor.

— Full mechanism, exact env vars, defaults, and size limits given. v0.12.0

02 **Two-Factor Sign-In (MFA)** NEW 95

Adds Two-Factor Sign-In requiring a code from an authenticator app at every sign-in, enabled via `ENABLE_MFA` or the Admin Settings > Authentication toggle (also reachable via Settings > Security). `MFA_ALLOW_OAUTH_BYPASS` and `MFA_ALLOW_TRUSTED_HEADER_BYPASS` exempt OAuth and trusted-header sign-ins, and `open-webui mfa reset <username>` issues a one-time recovery token for a locked-out user.

Enforce MFA for all users to require a one-time code from an authenticator app at every sign-in.

📍 In the admin panel, go to Settings › Security and enable `Two-Factor Sign-In (MFA)`.

Enforce MFA for all users while letting SSO users through without a second factor — useful when you run a mix of local and OAuth accounts.

```ini
ENABLE_MFA=true
MFA_ALLOW_OAUTH_BYPASS=true
```

Issue a one-time recovery token for an admin who has lost their authenticator app.

```
$ open-webui mfa reset <username>
```

— Names exact env vars and a runnable CLI command. v0.12.0

---

03 **Community hub for sharing presets, skills and tools** NEW

95

Adds a Community hub for sharing and importing model presets, prompts, skills (with all files, up to 15 MiB), tool source code, function source code (admins only), and full chat snapshots as Community posts. Adds `Workspace > Skills` as a dedicated area to create, import, and manage skills, with an Export Skills permission controlling who can share them, and Community share links on the Models, Prompts, Skills, Tools, and Functions pages and their per-item menus, reachable from `Settings > Admin > General`.

Share a skill (with all its files) to the Community hub so other teams can import and reuse it.

> 📍 In Open WebUI, go to Workspace › Skills, open the skill's menu, and click Export Skills. The skill and all its files (up to 15 MiB total) are published as a Community post.

Enable Community sharing links across Models, Prompts, Skills, Tools, and Functions pages for all users.

> 📍 In Open WebUI, go to Settings › Admin › General and enable community sharing. Share links then appear in the item menus on the Models, Prompts, Skills, Tools, and Functions pages.

— Exact UI paths, permission name, and size limit given. product docs

---

04 **Tool Search for large tool sets** NEW

95

Adds Tool Search for navigating large tool sets, configured via `ENABLE_TOOL_SEARCH`, `TOOL_SEARCH_DEFER_THRESHOLD`, `TOOL_SEARCH_ALWAYS_LOADED`, and

`TOOL_SEARCH_DEFER_BUILTIN_TOOLS` environment variables (and an Admin Settings > Interface toggle). This experimental mode keeps large tool definitions out of the prompt and exposes them via a `search_tools` tool on demand, preserving prompt-cache hits when many tools or MCP servers are active.

Enable Tool Search with a custom deferral threshold so only definitions longer than 800 characters are held back, and always keep your GitHub tools in every request.

```ini
ENABLE_TOOL_SEARCH=true
TOOL_SEARCH_DEFER_THRESHOLD=800
TOOL_SEARCH_ALWAYS_LOADED=github_*
TOOL_SEARCH_DEFER_BUILTIN_TOOLS=false
```

— Exact env vars plus a worked configuration example. v0.12.0

---

05 **Model Controls and parameter presets** NEW 75

Adds named Model Controls to the model editor (menus or sliders, each carrying custom parameters such as thinking effort), selectable per-model from a knobs button in chat input for users with the Allow Chat Controls and Allow Chat Params permissions, letting users pick admin-approved parameter presets directly in chat.

Give users a constrained set of model parameter presets (e.g. thinking effort levels) so they cannot set arbitrary parameters.

> 📍 In the admin panel, go to Models › [select model] › Model Controls and define the approved parameter presets users may choose from in chat.

— Names exact permissions and the UI entry point. v0.12.0

---

06 **Skills multi-file support and import/export** NEW 75

Adds multi-file support to workspace Skills, with a file tree, ZIP/JSON import and export, and a `/skills:create` chat command that saves a new skill directly to the workspace.

— Names an exact chat command and export formats. v0.12.0

---

07 **Knowledge base file browser redesign** IMPROVED 65

Adds a file browser with in-place preview for browsing a knowledge base's files and folders, redesigned with a folder/file tree, inline text editing of retrieved content, original-file preview, and search, filter, and sort controls.

Browse and preview knowledge base files without leaving the chat interface.

> 📍 In the chat sidebar, open the Knowledge Base section, then select File Browser to navigate folders and preview documents in place.

— Describes the redesign with a clear UI starting point. v0.12.0

08 **Version history for workspace items** NEW 65

Adds version history for saved configurations, Python scripts, and Markdown instruction sets, extended to workspace models, tools, functions, and skills — every save is versioned, browsable, diff-comparable, and promotable back to live from within each editor.

— Explains behavior but no exact flag or path named. v0.12.0

09 **Sharing mode for live chat replies** NEW 65

Adds a Sharing mode setting (Clone only or Allow replies) when sharing a chat, folder, or group, enabling all recipients to write in the same live conversation with real-time streaming, typing indicators, and per-user attribution.

— Describes the two modes and live behavior clearly. v0.12.0

THINNER COVERAGE BELOW

10 **Nested group support** NEW 55

Adds nested group support so administrators can assign a parent group; inner-group members inherit the parent's shared resources and permissions, and deleting a group promotes its subgroups to the parent.

— Explains inheritance and deletion behavior, no UI path given. v0.12.0

11 **Context compaction during tool-call runs** IMPROVED 40

Adds context compaction between tool-call rounds during long native-function-calling runs, preventing context overflow when many tool calls accumulate in a single reply.

— Mechanism named but no config or metric given. v0.12.0

12 **Built-in Python code editor for chat scripts** NEW 20

Adds a built-in code editor for Python scripts that run in the chat.

— Single sentence, no mechanism or location given. product docs

WAS THIS USEFUL? 👍 👎

◆ AI / LLM Gateways

BerriAI LiteLLM

Feedback ⓘ v1.104.3 NOTES ↗ RANK

LiteLLM is an AI gateway that calls over 100 model APIs in OpenAI format with cost tracking, guardrails, load balancing, and logging.

> LiteLLM shipped per-tenant control over OpenTelemetry prompt/response capture, a new lightweight `litellm-core` SDK distribution, and a decision model guardrail with matching Admin UI setup screen for decision-based routing.

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

01 Per-tenant content capture control in OTel v2

BREAKING

95

A new `capture_message_content` setting lets teams opt a specific team or key's OTel destination into prompt/response capture (`span_only`) or out of it, independent of the proxy's global setting; `/key/generate` and the team callback route reject any other value. When set to opt out, content attributes are stripped across every OTel vocabulary the proxy writes — `llm.input_messages.*` keys, MCP fields, OpenInference fields, and Langtrace fields — while span identity, model, token usage, cost, and tool definitions are preserved. Breaking change: a stored team or key callback entry with no `capture_message_content` value previously inherited whatever the first OTel callback in `callbacks` captured; it now follows the proxy's own exporter for the same backend, so existing entries that should keep receiving prompts and responses must explicitly set `capture_message_content: span_only` .

Opt a specific team's OTel destination into prompt and response capture while the global setting keeps other tenants metadata-only.

```yaml
capture_message_content: span_only
```

— Names exact config key, values, affected OTel vocabularies, and migration step.

product docs

02 Lightweight litellm-core SDK distribution

NEW

70

A new pip distribution, `litellm-core` , packages the Python SDK with fewer mandatory dependencies and without the bundled dashboard or gateway CLI entry points, while Python imports remain unchanged. It cannot be installed alongside `litellm` in the same environment since the two share the same Python files, so adopting it requires a fresh environment.

Install the lighter SDK distribution in a fresh environment when you only need the Python SDK and want to avoid pulling in the dashboard or gateway CLI dependencies.

```
$ pip install litellm-core
```

— Gives install command and environment constraint but little mechanism detail.

product docs

THINNER COVERAGE BELOW

03 Decision-based routing guardrail and setup UI NEW

40

Adds a decision model guardrail via `litellm.decisions` to the AI Gateway for routing and policy enforcement, alongside a decision auto-router setup screen in the Admin UI for configuring decision-based routing.

— Names the guardrail but gives no usage example or exact UI path. v1.104.3

NEW INTEGRATION Databricks Zerobus trace logging callback

NEW INTEGRATION xAI native batches and files support

NEW INTEGRATION Sail provider

└──▷ BREAKING ON UPGRADE

! A stored team or key callback entry with no `capture_message_content` value previously received whatever the first OTel callback in `callbacks` captured; it now follows the proxy's own exporter for the same backend. Set `capture_message_content: span_only` on each entry that should continue receiving prompts and responses.

! `litellm` and `litellm-core` cannot be installed in the same environment because they share the same Python files — installing `litellm-core` requires a fresh environment.

v1.105.0 NOTES ↗ STABLE

Adds fail-closed Redis rate limiting, SSO debug gating, Sentry PII scrubbing opt-in, MCP protocol version config, team budget email alerts, and new Rust gateway crates.

└──▷ GET THIS VERSION

```
$ git clone --branch v1.105.0 https://github.com/BerriAI/litellm.git
# already have the repo? check out this version:
$ git checkout v1.105.0
```

Adds `fail_closed_rate_limit_enforcement` to reject requests with HTTP 503 when Redis rate-limit counters are unreachable, preventing unenforced limits from

passing traffic.

Adds `ENABLE_SSO_DEBUG` flag (off by default) required to expose `/sso/debug` routes, reducing the SSO debug surface in production deployments.

Adds `SENTRY_SEND_DEFAULT_PII` opt-in flag alongside automatic scrubbing of PII and secrets inside object reprs and nested locals sent to Sentry.

Adds MCP protocol version configuration and capability discovery via the feat(mcp) surface, letting operators pin MCP versions per tool.

Adds email alerts triggered at configured percentages of a team member's spend budget.

+2 more

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

Feedback ⓘ snapshot-20261011 NOTES ↗

RANK

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

> LangSmith's latest snapshot adds an MCP tool for model pricing, deep-link support for SSO logins, model-access filtering on the LLM Gateway's model list, IAM role assumption for ECR pulls, and several breaking changes including a 180-day retention cap and renamed `repos:*` / `run-rules:*` permissions.

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

01 **Model access filtering on GET /v1/models** IMPROVED 80

LLM Gateway `GET /v1/models` now returns only models permitted by the calling credential's model access policies, suppressing models that would be rejected on use.

Check which models your credential is allowed to call before sending requests — the list now reflects your model access policies exactly.

```
$ curl -s https://<gateway-host>/v1/models \
  -H 'Authorization: Bearer <token>' | jq '.data[].id'
```

— Named endpoint with runnable curl example snapshot-20261011

02 **redirect_to parameter for SSO login** NEW 75

Adds `redirect_to` query parameter to the organization SSO login URL (`/sso/login/<slug>?redirect_to=<path>`) so deep links land on the target page after SAML sign-in instead of the workspace root.

Redirect users to a specific run after SAML sign-in instead of dropping them at the workspace root — useful for sharing deep links in incident response workflows.

```
$ https://app.langsmith.com/sso/login/<slug>?
redirect_to=/o/<org>/projects/p/<project>/r/<run>
```

— Exact endpoint and working example URL provided snapshot-20261011

03 **Permission renames: repos:* and run-rules:* retired** BREAKING 70

Built-in workspace roles no longer return the retired `repos:*` and `run-rules:*` permissions (renamed to `prompts:*` and `rules:*`); these are rejected when creating a custom role.

— Names exact permission strings on both sides of the rename snapshot-20261011

04 **Model Access policy configuration options** IMPROVED

60

Model Access policies gain a new Model configurations provider option that allows a workspace's saved model configurations (all or a chosen subset) so requests resolving to a configuration are permitted rather than blocked; new policies also include a generated policy name that can be kept or edited before creation.

— Both increments in same policy area, no exact UI path given snapshot-20261011

THINNER COVERAGE BELOW

05 **MCP pricing tool for model price maps** NEW 55

Adds `list_model_price_maps` tool to LangSmith MCP for searching and paginating model pricing rules in a workspace.

— Names the tool but no usage example or parameters given snapshot-20261011

06 **Agent Auth connection visibility for admins** NEW 55

Workspace administrators can view and inspect workspace-level Agent Auth connections — including authentication type, update details, and associated credential metadata — from the Integrations settings page.

— Names UI location but no field-level detail snapshot-20261011

07 **Trace retention capped at 180 days** BREAKING 55

Extended retention settings above 180 days are applied as 180 days for new traces ingested into LangSmith SaaS; new settings above 180 days are rejected outright.

— Clear before/after but no migration command snapshot-20261011

08 **Service key role assignment restrictions** BREAKING 55

Organization Operators can no longer grant the Organization Admin role to a service key, or change the roles of a service key that already holds it; a service key can no longer change its own roles.

— Clear restriction described, no API path given snapshot-20261011

09 **Raw TCP passthrough to allow-listed public IPs** NEW

50

Sandboxes now pass raw TCP connections to literal public IP addresses when the destination IP and port are explicitly allow-listed, supporting clients that dial a service's public IP directly.

— Describes mechanism but no config surface named `snapshot-20261011`

---

10 **IAM role assumption for ECR snapshot pulls** NEW 45

LangSmith Cloud can now assume tagged customer IAM roles for ECR repository discovery and snapshot image pulls.

— States capability without setup steps or role requirements `snapshot-20261011`

---

11 **Context-aware LangSmith Chat on tracing projects** IMPROVED 40

LangSmith Chat on a tracing project now reads active filters, view, time window, and selection, and names runs and threads by what was asked rather than by the serving app.

— Behavioural change described, no concrete command or setting `snapshot-20261011`

---

12 **Context Hub repository file limit raised to 1,500** IMPROVED 40

Context Hub repositories now support up to 1,500 files, allowing agents with larger skill collections to sync successfully.

— Clear numeric limit but no usage steps `snapshot-20261011`

└──▷ BREAKING ON UPGRADE

! Extended retention settings above 180 days are applied as 180 days for new traces ingested into LangSmith SaaS; new settings above 180 days are rejected outright.

! Built-in workspace roles no longer return the retired `repos:*` and `run-rules:*` permissions (renamed to `prompts:*` and `rules:*`); these are rejected when creating a custom role.

! Organization Operators can no longer grant the Organization Admin role to a service key, or change the roles of a service key that already holds it; a service key can no longer change its own roles.

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ snapshot-20261011 NOTES ↗

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

Adds Coding Agent Insights dashboards, SQL charts, Loop-on-dashboards, thread view, MCP-based log-to-dataset copying, and Windows DPAPI credential storage.

Adds `Dashboards > Coding agent insights` dashboard showing usage by user, API key, model, tool, and branch with cost breakdowns; requires data plane v2.13.0+.

UPDATED INTEGRATION Adds a daily Loop automation under Coding agent insights that finds recurring problems and inefficiencies across team sessions.

Adds SQL charts: build dashboard charts from a SQL query and switch between query text and a visual builder; available on Pro and Enterprise plans, requires data plane v2.1.1+.

Adds Loop-on-dashboards: open Loop from an individual dashboard or the chart editor to investigate data and build or refine charts in natural language.

UPDATED INTEGRATION Adds ability to copy logs into datasets using the Braintrust MCP server or Loop, preserving input, metadata, tags, and links to source logs; requires data plane v2.18.0+.

+10 more

NEW INTEGRATION pi 1.0 codemode tracing

UPDATED INTEGRATION Claude Code subagent tracing

UPDATED INTEGRATION OpenAI Responses API and Agents API span metadata: access_programs, computer use, browser auth

UPDATED INTEGRATION Claude Agent SDK span: verbatim_prompts

NEW INTEGRATION AgentScope TeamPipeline.reply_stream() tracing

└──▷ **BREAKING ON UPGRADE**

! On pi model call spans, `reasoning_tokens` is renamed to `completion_reasoning_tokens`, `cost` is renamed to `estimated_cost`, `metadata.provider_request` is replaced by top-level fields such as `temperature` and `tools`, and skill loads now use `tool_name: 'read'` with `tool_kind: 'skill'` — update any queries using the earlier field names.

! Spans no longer include `metadata.bt_daemon_version` — update any queries or automations that read this field.

WAS THIS USEFUL? 👍 👎

Arize Phoenix

Feedback ⓘ arize-phoenix-v20.21.0 NOTES ↗

RANK

Arize Phoenix is an open-source platform that monitors LLM application traces and evaluates their outputs.

> Phoenix v20.21.0 adds Mistral AI model pricing to its cost-tracking subsystem and gives ingestion-created projects a random gradient for visual distinction.

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

01 **Mistral AI pricing in cost tracking** NEW 30

The cost-tracking subsystem now supports Mistral AI model pricing, allowing cost calculations for Mistral-based traces.

— Names the provider but no mechanism or scope detail arize-phoenix-v20.21.0

02 **Random gradients for ingestion-created projects** IMPROVED 25

Projects created via ingestion are now assigned a random gradient.

— Bare description, no further mechanism given arize-phoenix-v20.21.0

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Open-Weight Models

DeepSeek V3.2 Exp

Feedback ⓘ price increase · Oct 10, 2026

MODEL ↗

DeepSeek V3.2 Exp is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v3.2): $0.259 → $0.259 in, $0.42 → $0.80 out per million tokens (+90%), first seen 10 Oct 2026.

WAS THIS USEFUL? 👍 👎

Z.ai GLM 4.6

Feedback ⓘ price increase · Oct 10, 2026 MODEL ↗

GLM 4.6 is a chat model from Z.ai.

Price increase on OpenRouter (z-ai/glm-4.6): $0.43 → $0.50 in, $1.75 → $2.00 out per million tokens (+16%), first seen 10 Oct 2026.

WAS THIS USEFUL? 👍 👎

Z.ai GLM 5.2

Feedback ⓘ price cut · Oct 10, 2026 MODEL ↗

GLM 5.2 is a chat model from Z.ai.

Price cut on OpenRouter (z-ai/glm-5.2): $0.41 → $0.06 in, $3.99 → $7.00 out per million tokens (-85%), first seen 10 Oct 2026.

WAS THIS USEFUL? 👍 👎

Google Gemma 4 26B A4B

Feedback ⓘ

price cut · Oct 10, 2026 MODEL ↗

Gemma 4 26B A4B is a multimodal model from Google.

Price cut on OpenRouter (google/gemma-4-26b-a4b-it): $0.09 → $0.0675 in, $0.30 → $0.225 out per million tokens (-25%), first seen 10 Oct 2026.

WAS THIS USEFUL? 👍 👎

Moonshot AI Kimi K2.5

Feedback ⓘ

price increase · Oct 10, 2026 MODEL ↗

Kimi K2.5 is a multimodal model from Moonshot AI.

Price increase on OpenRouter (moonshotai/kimi-k2.5): $0.45 → $0.50 in, $2.25 → $2.50 out per million tokens (+11%), first seen 10 Oct 2026.

WAS THIS USEFUL? 👍 👎

NVIDIA Nemotron 3.5 Lightning

Feedback ⓘ

price increase · Oct 10, 2026 MODEL ↗

Nemotron 3.5 Lightning is a chat model from NVIDIA.

Price increase on OpenRouter (nvidia/nemotron-3.5-lightning): $0.049 → $0.07 in, $0.14 → $0.20 out per million tokens (+43%), first seen 10 Oct 2026.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · October 11, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue