THE AI TOOLCHAIN

NO. 055

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED OCTOBER 10, 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 29 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 2 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. Both OpenAI SDKs can now authenticate from Kubernetes, Azure, GCP or X.509 with no API key in the environment at all. OpenCompass adds an LLM-judge and a LaTeX-

vim ? help

aware math verifier so free-text answers stop needing exact-match hacks. Cline runs subagents in parallel and on cron schedules.

GOVERN

Call OpenAI from a pod with no API key to leak

Both SDKs take workload identity from Kubernetes, Azure, GCP and X.509 mTLS (plus custom JWT in Python), so the credential is minted by the platform per workload instead of sitting in a secret store and rotated by hand. The TypeScript SDK also verifies webhook signatures in-library, which removes the usual hand-rolled HMAC check.

OpenAI Python SDK · OpenAI TypeScript SDK

EVALUATE

Score free-text and math answers without writing an exact-match regex

MATHVerifyEvaluator compares LaTeX answers by equivalence rather than string match and GenericLLMEvaluator runs a model as judge, while CascadeEvaluator falls back to the expensive judge only when cheap scoring is ambiguous. CustomDataset/ChatMLDataset paths mean a new benchmark no longer needs a Dataset subclass before it can be run.

OpenCompass

BUILD

Fan a coding task out to parallel agents, or let one run at 3am

Cline's `spawn_agent` and `/team` split work across subagents instead of serializing it in one context window, and `cline schedule` plus the SDK automation API put runs on cron without a human kicking them off. Letta Code drives agents across multiple machines and swaps between cloud and local backends.

Cline · Letta Code

DEPLOY

Serve MoE and 4-bit checkpoints without a bespoke conversion script

compressed-tensors adds a MoEExpertPacker for repacking linearized MoE checkpoints and a QuantizationConfig/ModelCompressor API covering MXFP4, MXFP8 and NVFP4; AutoRound's `svdquant_omni` export loads SVDQuant NVFP4 checkpoints straight into vLLM-Omni. The glue code between a quantizer's output and a serving engine's expected layout is the part that usually gets rewritten per model.

compressed-tensors

BUILD

◆ AI Coding Agents

yetone magpie

Feedback ⓘ v0.1.1161 CODE ↗ RANK

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 four releases, magpie added plugin-defined AI agents with pluggable config formats, Codex Fast tier routing plus retry budgets for busy routing groups, Claude Code binary auto-download with SHA-256 verification, custom provider setup from the TUI, and a long tail of provider-specific compatibility fixes (Kimi sampling parameter stripping, Claude Opus reasoning visibility, Gemini thinking-level retries, MCP auth and health-check changes).

└──▷ **WHAT SHIPPED · 27 FEATURES** most completely described first ? what's the number?

01 **Plugin-defined agents via `magpie.agent`** NEW 91

Plugins can register their own AI coding agents by declaring a `magpie.agent` field in their `package.json` pointing to a JS module that exports an `agent` object (with `id`, `config`, `model`, `prefix`) and a `connect` function returning gateway credentials, letting plugins wire external agents (e.g. Aider) through magpie's gateway without custom write logic. Agent config files can be JSON/JSONC, YAML/YML, or `.env`, selected via the `agent.format` field or file extension, as enumerated in the `agentplug` package.

Wire an external agent (e.g. Aider) through magpie's gateway by declaring a plugin agent module in your plugin's `package.json`.

```json
{"name": "my-aider-plugin", "magpie": {"agent": "./agent.js"}}
```

Define the agent module that tells magpie which config file to edit and what credentials to inject for the external agent.

```json
export const agent = {
  id: "aider",
  name: "Aider",
  config: "~/.aider.conf.yml",
  model: "model",
  prefix: "openai/"
};
export function connect({ gateway, model, models }) {
  return { "openai-api-base": gateway.v1, "openai-api-key":
gateway.key };
}
```

— Full mechanism, runnable code example, named config fields v0.1.1161

02 **Claude Code binary auto-download** NEW 77

Downloads Anthropic's own Claude Code build (from `https://downloads.claude.ai/claude-code-releases`) for servers or containers without it, verified against the release `manifest.json` SHA-256 and size before use, stored in magpie's cache folder under `claude-code/`.

— Full mechanism and paths named, happens automatically v0.1.1158

03 **Codex Fast tier exposure for remote instances** NEW 76

Exposes `service_tiers` on each model object in the models API response when a remote magpie instance is queried, carrying the Codex tiers that instance offers, so a client can request Codex Fast on a remote magpie's model (#1234).

— Named API field and endpoint behavior with issue reference v0.1.1159

04 **Per-agent token-optimization tuning** NEW 76

Adds per-agent token-optimization tuning (`compact`, `cache_ttl`, `subagent_cache_ttl`) that replays the user's own call history to recommend the lowest-token settings, applied on a click and undone on another.

— Named settings and mechanism, UI action described v0.1.1158

05 **Kimi sampling parameter rewriting** IMPROVED 70

Rewrites (strips) locked sampling parameters `temperature` and `top_p` from requests to Kimi models that reject them with a 400, covering `kimi-for-coding*`, `k3*`, `kimi-k2.5` and later on `api.moonshot.ai`, `api.moonshot.cn`, `api.kimi.com`, and `api.kimi.ai` (#1197).

— Exhaustively named models/hosts but automatic fix, no user action v0.1.1159

06 **Retry budget for busy routing groups** NEW 68

Adds a retry budget (`Patience` , 1 minute by default) for routing groups whose every member is busy, re-asking each member within that window instead of terminating the turn on the last member's error (#1418).

— Named setting and default value, no manual config path shown v0.1.1159

07 **`magpie group copy` subcommand** NEW 68

Adds `magpie group copy` subcommand to duplicate a routing group from the Routing page menu.

Duplicate an existing routing group to use as a starting point for a new configuration, without rebuilding it from scratch.

```
$ magpie group copy
```

— Runnable CLI command with exact name given v0.1.1158

08 **Search API off mode and multi-key failover** NEW 67

Adds `Searcher = "off"` to settings so a model's web search bypasses the provider and goes only to configured Search APIs. Search APIs now accept several keys; when one key runs out of credit, the search is handed to the next key automatically.

Force all web search through your own Search APIs only, ignoring any provider-side search capability.

```env
st.Searcher = "off"
```

— Named config flag with runnable example and failover mechanism v0.1.1160

09 **Proxy forwarding to plugin Node programs** NEW 62

Passes magpie's proxy to Node programs started by a plugin via `NODE_USE_ENV_PROXY=1` , so Node's `fetch` honours `HTTPS_PROXY` .

— Named env var, clear mechanism, no command needed v0.1.1159

10 **Custom provider setup from TUI Providers page** NEW 62

Adds custom provider support from the TUI Providers page: add any OpenAI, Responses, Anthropic, or Gemini compatible provider by entering its base URL.

— Clear UI navigation and compatible types, no exact field name v0.1.1158

11 **Snow CLI model controls on Agents page** NEW 60

Snow CLI's light model (`basicModel`) can now be selected independently from its main model on the Agents page. Snow CLI's thinking level can now be set from the Agents page, using the levels available to magpie's model. Snow CLI's window, output, and image support now follow magpie's model catalog until the user overrides them in Snow's own settings.

— Names field and UI location but no exact command v0.1.1160

12 **Per-URL MCP health-check timeout budget** IMPROVED 60

Gives each MCP server health-check step its own 60-second budget per URL, reports which step went unanswered and for how long, and removes the per-session timeout (#1467).

— Concrete timing and behavior but no config shown v0.1.1159

13 **Claude Opus reasoning visibility** IMPROVED 60

Forwards `thinking.display "summarized"` when a Chat or Responses client asks Claude Opus 4.7 or later to reason, so the reasoning is visible rather than omitted (#1485).

— Named field and model version but automatic behavior v0.1.1159

THINNER COVERAGE BELOW

14 **Cached prompt reuse for Codex/ChatGPT-API backends** IMPROVED 57

Reuses cached prompts for tool-enabled conversations in the `codex` and `chatgpt-api` backends, sending session state via `session-id` / `thread-id` headers as Codex 0.162 does.

— Names backends and headers but automatic, no user action v0.1.1161

15 **Claude Desktop Microsoft Store package detection** IMPROVED 53

Finds Claude Desktop installed from the Microsoft Store (`AnthropicPBC.Claude_*` package) alongside the setup.exe `Claude_*` package when locating desktop folders.

— Named package IDs but passive detection only v0.1.1158

16 **MCP sign-in via client_secret_basic** NEW 50

Supports sign-in to MCP servers whose token endpoint requires credentials only via `client_secret_basic` , regardless of what the server's metadata lists first (#1490).

— Names auth method but no setup steps v0.1.1159

17 **New provider and agent visibility: Cursor Private Inference, AiHubMix**

NEW 46

Lists Cursor Private Inference on the Agents page when installed in a user-chosen folder on Windows or kept as its AppImage on Linux (#1254). Surfaces AiHubMix beside OpenRouter in the homepage model list.

— Names both surfaces but each addition thin individually v0.1.1159

18 **Expanded model testing coverage** NEW 45

Allows model tests to be sent as Codex or Claude Code would send them, enabling relay testing for providers that only serve those clients. Enables testing models from an Antigravity or Gemini CLI sign-in through the gateway's translator.

— Names clients covered but no concrete steps given v0.1.1158

19 **Copy MCP servers to project file** NEW 40

MCP servers, or a selected subset, can be copied as an agent's project file has them, for pasting into a project magpie doesn't manage.

— Describes action but no exact UI path or command v0.1.1160

20 **'Tuned for you' context cost view** NEW 40

The Tuned for you context view draws each setting's cost as a curve with waste shaded and shows how much more the current window setting allows.

— Describes visualization but no navigation path v0.1.1160

21 **Gemini thinking-level retry** IMPROVED 39

Remembers when Gemini rejects a thinking level and retries at the model's lowest available level rather than disabling thoughts for the provider globally.

— Behavior described, no named parameter or config v0.1.1159

22 **Codex idle-timeout handling** IMPROVED 37

Improves Codex idle-timeout handling so magpie delivers the reply event within the idle window instead of reconnecting on 'idle timeout waiting for SSE'.

— Explains before/after but no user action v0.1.1161

23 **Repetitive reply loop detection in gateway** NEW 35

The gateway now detects and terminates repetitive reply loops (e.g. 'mcp mcp mcp ...'), covering both word-level and line-level loops, ending runaway responses automatically.

— Describes behavior but no config or trigger to act on v0.1.1161

24 **WorkBuddy plugin enterprise account support** IMPROVED

35

WorkBuddy plugin updated to 0.1.13, adding support for enterprise accounts to list and use enterprise models.

— Names version but limited usage detail v0.1.1160

25 **Ante model context-window listing** IMPROVED

35

Ante now lists every magpie model whose reply limit reaches its context window, reported one token under it.

— Specific number given but no actionable step v0.1.1160

26 **Root-level union type handling for OpenAI API** IMPROVED

30

Handles root-level union types rejected by OpenAI's API by re-submitting the request in a folded form.

— Mechanism named but no concrete trigger or config v0.1.1161

27 **Magpie web link printing behind reverse proxy** IMPROVED

25

Magpie web behind a reverse proxy now prints its link through the proxy, including the key.

— Thin description, no config or command shown v0.1.1160

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Inference Providers & GPU Clouds

Modal

Feedback ⓘ changelog · Oct 10, 2026 CHANGELOG ↗ RANK

Bring your own code, and run CPU, GPU, and data-intensive compute at scale. The serverless platform for AI and data teams.

> Modal 1.6.1 adds CLI observability for Modal Endpoints and a new container-count metric across function and server stats.

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

01 Modal Endpoint CLI observability commands NEW 75

Three new CLI commands for inspecting deployed Modal Endpoints: `modal endpoint info` displays basic metadata about an Endpoint, `modal endpoint stats` reports aggregate performance metrics, and `modal endpoint logs` fetches or streams logs for an Endpoint.

Stream live logs from a deployed Modal Endpoint to debug latency or errors in real time.

```
$ modal endpoint logs <endpoint-name>
```

Check aggregate performance metrics for an endpoint, including request counts and latency, from the CLI.

```
$ modal endpoint stats <endpoint-name>
```

— Names three exact commands with runnable examples changelog-20261010-113aba36

02 container_total_count metric on stats NEW 60

Adds a `container_total_count` field to `modal function stats`, `modal server stats`, Function.stats(), and Server.stats(), reporting the number of containers live at any point in the interval.

— Names exact field and four surfaces but no usage example changelog-20261010-113aba36

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · October 10, 2026

Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe Feedback on this issue