THE AI TOOLCHAIN

NO. 042

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED SEPTEMBER 30, 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 392 changes

Do you prefer **scroll**? 👍 👎

▸ // CONTENTS SHOW 61 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. Qdrant's Constella preview swaps the query encoder without re-embedding a collection, so you can retune h speed against accuracy on existing data. Pinecone's namespace

vim ? help

aliases repoint reads to a freshly built namespace atomically. Axolotl adds GGUF export from the training config, ending the separate convert step before local serving.

> BUILD
>
> **Change your retrieval quality/latency tradeoff without rebuilding the index**
>
> Constella's hot-swappable query encoders decouple the model you search with from the vectors already stored, so tuning no longer means a full re-embed run; Pinecone's movable namespace pointers give the same escape hatch at the data level, letting a rebuilt namespace go live with no read downtime. Haystack also trims per-query overhead in `SentenceWindowRetriever`.
>
> Qdrant · Pinecone · Haystack

> BUILD
>
> **Get a quantized GGUF out of a fine-tune without leaving the YAML**
>
> Axolotl v0.20.0 drives GGUF export, Ringmaster context parallelism, native NVFP4 LoRA and expert parallelism from the same config file, so the usual hand-run conversion and sharding scripts between training and local inference disappear. TileLang v0.1.15 adds a Huawei Ascend 950 NPU backend with CMake flags per backend, for teams writing kernels off NVIDIA.
>
> Axolotl · TileLang

> OPERATE
>
> **Stop an agent run dying on a provider refusal or a dropped lease**
>
> Mastra 1.72.0 adds a cybersecurity-refusal retry handler and lease-fenced background tasks, so a model declining a security-adjacent request or a worker losing its lease no longer strands a workflow mid-flight; live channel resolution means changing a channel target doesn't require a redeploy. Cline handles Anthropic refusals and content filters with a fallback across both the extension and SDK.
>
> Mastra

BUILD

◆ AI Coding Agents

yetone magpie

Feedback ⓘ v0.1.515 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 biggest window addition is a full plugin ecosystem — CLI management, an in-app marketplace backed by `GET /api/plugins/market`, and multi-account plugin providers — alongside encrypted WebDAV/S3 config sync, custom per-model price overrides, a Docker server image, and new agent integrations for OpenChamber, OmO, Droid, and Antigravity, plus MCP server support extended to eight more agents and dozens of smaller routing, display, and reliability fixes.

└──▷ **WHAT SHIPPED · 41 FEATURES** most completely described first ? what's the number?

01 Docker server image with LAN sharing NEW 98

`docker build` produces a distroless, nonroot server image with magpie's files in a volume at `/config`, exposing the agent gateway on port 3425 and the browser UI on port 3430. `magpie web --addr 0.0.0.0:3430 --no-open` runs the browser UI inside the container, with `MAGPIE_WEB_KEY` keeping a stable key across restarts. `"lan": true, "lanKey": "sk-magpie-…"` keys in `/config/magpie/settings.json` (or Settings → Share on local network) gate access from outside the container behind a required API key before exposing ports beyond 127.0.0.1.

Run magpie as a hardened Docker server with LAN sharing locked behind a key — safe to expose beyond loopback once the key is set.

```
$ docker build -t magpie .
docker run -d --name magpie -p 127.0.0.1:3425:3425 -p
127.0.0.1:3430:3430 -v magpie-config:/config magpie
```

Enable LAN access with a required API key so agents on other machines can reach the gateway without opening it to unauthenticated traffic.

```json
{"lan": true, "lanKey": "sk-magpie-yourSecretKeyHere"}
```

— Ports, flags, and config keys are all runnable. product docs

02 Custom per-model and provider price overrides NEW

95

`magpie model price <provider/model> <in>,<out>,<cache-read>,<cache-write>` sets a custom USD-per-million-token cost for any model or provider wildcard such as `relay-a/*`, overriding models.dev list prices for resellers and unlisted models. `magpie model price <provider/model> --reset` removes a user-set price and falls back to the lookup chain (per-model → `<provider>/*` → provider catalogue → models.dev), and `magpie model prices` lists every model or pattern that has been explicitly priced.

Set accurate per-token pricing for a relay provider that resells GPT models at a different rate than models.dev lists.

```
$ magpie model price relay-a/gpt-5.5 0.12,0.60,0.01,0.15
```

Apply a single flat price to every model under a provider with a wildcard, covering unlisted or custom models in bulk.

```
$ magpie model price relay-a/* 0.12,0.60,0.01,0.15
```

Audit all custom price overrides you have set across providers before a billing review.

```
$ magpie model prices
```

Correct the cost ledger for a relay provider that resells gpt-5.5 at a discount — set actual input/output/cache-read/cache-write prices in USD per million tokens.

```
$ magpie model price relay-a/gpt-5.5 0.12,0.60,0.01,0.15
```

— Full command set with reset flag and lookup chain given. v0.1.504

03 **Plugin install, login, and marketplace management** NEW

93

`magpie plugin add <npm-package-or-path>` installs OpenCode-compatible auth plugins; `magpie plugin login <plugin>`, `magpie plugin logout <plugin>`, `magpie plugin off <plugin>`, `magpie plugin on <plugin>`, `magpie plugin rm <plugin>`, and `magpie plugin update` manage installed plugins and sign-in state, also reachable via Settings → Plugins UI. Metadata is stored in `~/.config/magpie/plugins.json`, packages in `~/.config/magpie/plugins/`, auth in `~/.config/magpie/plugin-auth.json` (mode 0600), and the Bun runtime in `~/.cache/magpie/bun/`. A `GET /api/plugins/market` endpoint returns the

community plugin registry with live npm metadata, overridable via `MAGPIE_PLUGIN_MARKET` (set to `off` to force the bundled copy), backing a Plugins tab with Discover and Installed views.

Install a Gemini auth plugin so agents can call Gemini models through your OpenCode plan without a direct API key.

```
$ magpie plugin add opencode-gemini-auth
magpie plugin login google-plugin
```

Add a Gemini auth plugin so magpie can sign in to a Gemini subscription without a separate API key.

```
$ magpie plugin add opencode-gemini-auth
magpie plugin login google-plugin
```

Query the plugin market API directly to see what provider plugins are available, including community listings and live npm metadata.

```
$ curl http://localhost:<port>/api/plugins/market
```

Force magpie to use only the built-in bundled plugin registry without reaching out to GitHub — useful in air-gapped environments.

```
$ MAGPIE_PLUGIN_MARKET=off magpie
```

Point magpie at a self-hosted plugin registry instead of the default community repo.

```
$
MAGPIE_PLUGIN_MARKET=https://registry.internal.example.com/plugin
s/registry.json magpie
```

— Names every subcommand, storage path, and endpoint. v0.1.489

04 **OmO agent integration** NEW 93

Adds OmO (`omo-ai` / `oh-my-openagent`) as a new managed agent: magpie sets `defaultProvider` , `defaultModel` , and `defaultThinkingLevel` in `~/.omo/agent/settings.json` and registers itself as the `magpie` provider in `~/.omo/agent/models.json` , mirroring the Pi integration without touching `~/.pi` .

Supports OmO's MCP server config at `~/.omo/agent/mcp.json` (Pi-native format), agent instructions at `~/.omo/agent/AGENTS.md`, and skills at `~/.omo/agent/skills`; `OMO_CODING_AGENT_DIR` and `SENPI_CODING_AGENT_DIR` environment variables relocate OmO's agent folder away from the default `~/.omo/agent`.

— Config paths and env vars fully named. v0.1.454

05 **Encrypted config sync via WebDAV and S3** NEW 92

`magpie webdav on <url> user=<user>` enables encrypted WebDAV sync (supporting 坚果云, Nextcloud, Synology), keeping all machines in sync every 3 minutes. `magpie s3 on <s3-url>` with `endpoint`, `region`, `path-style`, and `access-key-id` options enables encrypted S3-compatible sync (AWS S3, Cloudflare R2, Backblaze B2, MinIO, Garage); `magpie s3` shows status, `magpie s3 now` forces a sync, and `magpie s3 off` disables it. Both are also configurable via Settings → Sync and backup UI.

Keep magpie config encrypted and in sync across machines using a Cloudflare R2 bucket.

```
$ magpie s3 on s3://my-bucket/magpie
endpoint=https://<account>.r2.cloudflarestorage.com access-key-
id=<key>
```

Keep settings encrypted and in sync across machines using Cloudflare R2.

```
$ magpie s3 on s3://my-bucket/magpie
endpoint=https://<account>.r2.cloudflarestorage.com access-key-
id=<key>
```

— Commands, flags, and supported backends all named. product docs

06 **Kimi Code 2.x support and per-model thinking levels** NEW 90

Adds Kimi Code (2.x) support, reading settings from `~/.kimi-code/config.toml` — distinct from the legacy `~/.kimi/config.toml` — so models selected in magpie route correctly to Kimi Code's `/model` endpoint. Also writes `support_efforts` and `default_effort` fields into that config for each thinking-capable model, so Kimi Code's thinking picker presents actual effort levels (`low`, `high`, `max`) instead of a plain on/off toggle.

After upgrading, inspect the generated Kimi Code config to confirm thinking-level fields are populated for your models.

```toml
[models."<provider>/<model-id>"]
capabilities = ["thinking"]
support_efforts = ["low", "high", "max"]
default_effort = "high"
```

— Config paths and fields named with a config example. v0.1.515 v0.1.470

07 **Automatic [1m] suffix for 1M-context Claude models** IMPROVED

90

Automatically appends `[1m]` to any Claude model reference in `model`, `env.ANTHROPIC_MODEL`, `env.ANTHROPIC_SMALL_FAST_MODEL`, and tier fields of `~/.claude/settings.json` whose context window is 1 million tokens or more, so Claude Code 2.1.284 and earlier treat it as a 1M-context model instead of defaulting to 200K and triggering repeated autocompaction.

— Exact fields, file, and version threshold all named. v0.1.511

08 **MCP server integration for Hermes, omp, Qoder, Kimi Code, Devin, Grok Build, Cline, and Command Code**

NEW

90

Adds MCP server integration for Hermes via `mcp_servers` in its `config.yaml` (streamable HTTP and SSE); for omp and Qoder via their `mcp.json` with a `type` field accepting `http` or `sse`; for Kimi Code (Python kimi-cli and new versions) via `mcp.json` using a `transport` field of `http` or `sse`; for Devin via `mcp_config.json`, writing a `transport` field (`stdio`, `http`, or `sse`) on every entry; for Grok Build via `[mcp_servers.<name>]` tables in `config.toml` (no SSE support); for Cline CLI via `cline_mcp_settings.json` where transport is a nested object with `type` (`stdio`, `streamableHttp`, or `sse`); and for Command Code via `mcp.json` with a `transport` field (`stdio` or `http`) and an `enabled` flag.

— Every config file and transport field named per agent. v0.1.496

09 **X-Magpie-Account header for pinning gateway requests** NEW

88

Adds X-Magpie-Account request header to pin a gateway request to one specific account within a multi-account subscription, bypassing round-robin selection. Returns HTTP 429 if the named account is unavailable, without falling back to others.

Pin a probe or test request to one specific account (by email or account ID) in a subscription that holds several, so you can isolate behavior per account.

```
$ curl -X POST https://<magpie-host>/v1/responses \
  -H 'X-Magpie-Account: me@example.com' \
  -H 'Content-Type: application/json' \
  -d '{"model":"gpt-5.5","input":"Hello"}'
```

— Header, behaviour, and runnable curl example given. v0.1.509

10 **OpenChamber agent support** NEW 88

Adds an `openChamber` agent row reading settings from `$OPENCHAMBER_DATA_DIR` (defaulting to `~/.config/openchamber`) via `preferences.json` and `settings.json`, and respecting `$OPENCODE_CONFIG_DIR` for the bundled OpenCode instance. Applies a per-session model-priority chain — project `defaultModel` in `settings.json` projects[], then profile `defaultModel` in `preferences.json`, then the agent's own model, then OpenCode's model — and supports `smallModelOverride` (when `smallModelUseDefault` is false) for titles and commit messages, falling back to OpenCode's `small_model`.

— Full priority chain and env vars named, no CLI shown. v0.1.502

11 **Pi agent directory relocation and session visibility** NEW 86

Adds `PI_CODING_AGENT_DIR` environment variable to relocate Pi's agent folder (storing `settings.json` and `models.json`), with `~` `-expansion` support; relative paths and foreign-machine variables are ignored. Also exposes Pi's local sessions under Usage › Sessions in the GUI.

Relocate Pi's agent config to a custom dotfiles directory so settings.json and models.json live outside the default ~/.pi/agent path.

```
$ export PI_CODING_AGENT_DIR=~/dotfiles/pi
```

— Env var with a runnable export example. v0.1.490

12 **Baidu Qianfan Token Plan provider** NEW 86

Adds `baidu-qianfan` as a new provider (also registered as `qianfan-token-plan`), supporting personal (个人版), enterprise (企业版), and pay-as-you-go Baidu Qianfan Token Plans, each with Chat Completions, Responses, and Anthropic Messages endpoints; add via `magpie provider add baidu-qianfan`.

Add Baidu Qianfan Token Plan as a provider to route requests through its Chat Completions, Responses, and Anthropic Messages endpoints.

```
$ magpie provider add baidu-qianfan
```

— Runnable add command plus endpoint list given. product docs

13 **Native API routing preserves Claude thinking signatures**

IMPROVED 85

Adds `api` and `baseUrl` fields to `ompModel` in `models.yml` so each model is routed to its provider's native API — Anthropic Messages at `/v1/messages`, OpenAI Responses at `baseUrl/responses`, or Chat — instead of being written as `openai-completions`. Anthropic-served models route through the gateway's `/v1/messages` endpoint, preserving Claude extended thinking signatures and the `context_management`, `thinking.display`, and `output_config.effort` fields across tool turns.

— Endpoints and fields named, no user-facing toggle. v0.1.513

14 **Bulk hide/show for agent model lists** NEW 85

Adds per-agent model visibility control: a line under each agent's name on the Agents page shows Showing 5 / 32 models or All 41 models and opens a picker to hide or restore individual models, stored in `settings.HiddenModels`. Adds a Hide all button to the agent model list panel to hide every visible model at once (excluding the model currently in use), alongside the renamed Show all button (formerly Show all again).

— Setting name and exact UI counters both given. v0.1.463 v0.1.460

15 **Multiple accounts per plugin provider** NEW 80

Plugin providers can now hold multiple signed-in accounts, stored in `plugin-auth.json` under the provider's id for the first account and `id#<slot>` for subsequent ones, matching how built-in subscriptions manage accounts. Adds an Accounts field to the Provider struct in `internal/plugin/providers.go` exposing each account's Key, Type, `AccountID`, and Hint, and isolates account scope per request via `AsyncLocalStorage` added to `internal/plugin/host.js` so `getAuth`, `client.auth.set`, and the loader always use the correct account.

— Internal struct and field named, no CLI command shown. v0.1.500

16 **Antigravity (agy) Library target** NEW 80

Adds `agy` as a named Library target, reading rules from `~/.gemini/config/GEMINI.md`, MCP servers from `~/.gemini/config/mcp_config.json`, and skills from `~/.gemini/config/skills`. Adds `fmtAntigravity` MCP format support, encoding

remote servers with `serverUrl` (plus `type: sse` for SSE transports) and decoding both `serverUrl` and fallback `url` fields from Antigravity's `mcp_config.json`.

— Config paths named, no CLI command shown. v0.1.459

17 **Codex Bank Reset auto-spend** NEW 78

Adds `CodexAutoReset` setting (stored in `settings.CodexAutoReset`) letting a named Codex account automatically spend one of its Bank Reset credits when its weekly window is exhausted, off by default. Persists auto-reset state to `codex-autoreset.json` so a process restart does not spend an additional reset.

— Setting and persistence file named, no CLI shown. v0.1.507

18 **Manual routing and reasoning-level declarations for groups** NEW 78

Adds a Manual routing mode for groups, letting each request go to a hand-picked model selected on the group's card. Also adds a Levels field on routing groups so a group can explicitly declare which reasoning levels (`none`, `low`, `medium`, `high`, `xhigh`, `max`) it advertises, overriding the intersection of what its members individually support.

— Field and enum values named, UI-based action only. v0.1.498 v0.1.477

19 **Skills support extended to 13 more agents** NEW 75

Adds skills support for 13 more agents — including `kimi`, `goose`, `cindy`, `commandcode`, `devin`, `hermes`, `droid`, `cline`, `qoder`, `grok`, `workbuddy`, `fx`, and `devin` — routing each to its own skills folder or the shared `~/.agents/skills` directory. Agents whose home path (e.g. `~/.dsh`) is occupied by a file are now silently excluded from the library target list and skipped during writes rather than returning an 'open ... not a directory' error.

— Every agent named, but no per-agent path given. v0.1.469

20 **Usage and allowance display parity for Claude and plugin accounts** IMPROVED 73

Claude subscription usage-limit resets, queried via `/api/oauth/usage?cedar_ember=1&skip_spend=1`, now surface on each Claude account card, usable manually or automatically when the week's quota is exhausted. Plugin accounts now report allowance through `auth.usage`, surfacing `plan`, `user`, `until`, `renew`, `balance`, and `reset` fields, matching built-in account display parity.

— Endpoint and field names given, no UI path. v0.1.514

21 **Droid (Factory) agent integration** NEW 73

Adds Droid (Factory) integration with config at `~/.factory/settings.json` (overridable via `$FACTORY_HOME_OVERRIDE`), supporting magpie's models as BYOK `customModels` in Droid's `/model`.

— Config path and override variable named. product docs

22 **Provider duplication, account pausing, and per-key Balance URLs**

NEW 73

Adds `{key}` placeholder support in the Balance URL so each API key is named individually when checking balances. Adds provider duplication via `AddCopy`, cloning a provider's keys, balance token, and routing alongside the original, and adds the ability to pause a signed-in account on a provider without removing it.

— Mechanism named, no exact UI path given. v0.1.455

23 **Menu bar multi-subscription usage display** IMPROVED 70

Settings → Usage in the menu bar now displays any number of ticked subscriptions side by side, each rendered as its provider logo with two quota windows stacked (e.g. `42%` over `18%`) on macOS, or as pipe-separated text (`42% · 18% | 10% · 5%`) on other platforms.

— UI location given, viewing only, no further action. v0.1.499

24 **Claude Desktop 3p MCP sync and proxy-aware update checks**

IMPROVED 68

Writes library MCP servers to `Claude-3p/claude_desktop_config.json` so Claude Desktop's 3p gateway mode picks them up alongside the standard `Claude/claude_desktop_config.json`. Update checks now route through magpie's configured proxy, and a failed direct update check is retried immediately when a proxy is saved on the Settings page rather than waiting up to six hours.

— File path named, no user-facing toggle. v0.1.474

25 **Model maker logos in agent model lists** NEW 68

Adds a `logo` field (JSON: `logo`) to `agentModelJSON` so each model row in an agent's model list carries its maker's logo, resolved by model family or provider. Exports ModelIcon(model string) in `internal/agent/icons.go` to look up a maker icon by model ID or name alone (returning `""` when unrecognised), and adds recognition for `mimo` / `xiaomi` (icon: `mimocode`) and `step` (icon: `stepfun-color`) model families.

— Fields and function named, purely visual change. v0.1.466

26 **Codex-native model unreachable error message** IMPROVED

65

When a request targets one of Codex's own models (e.g. `gpt-5.5`) and OpenAI is unreachable, the gateway now returns a 502 with a human-readable message identifying the model as Codex-native and prompting the user to pick one of magpie's models instead of surfacing a raw dial error.

— Describes new behaviour, no config to change. v0.1.501

27 **Claude account removal in Providers UI** NEW

65

Adds a Remove button in the Providers UI for Claude accounts that are saved in magpie but signed out (e.g. banned accounts), letting practitioners purge magpie's stored credentials without touching Claude Code's own files or keychain.

— Exact UI button and effect named. v0.1.487

28 **Output-length capping at model context window** IMPROVED

63

Caps `maxTokens` sent to omp, Pi, Cline, and dsh agents at the model's `contextWindow` when the model's listed output limit exceeds its context window, preventing agents from requesting replies longer than the entire context.

— Mechanism stated but no config surface to act on. v0.1.514

29 **Gateway compatibility fixes for Copilot, Codex Responses, and gpt-6.x refusals**

IMPROVED

63

Enables the Copilot backend to auto-negotiate the API version VS Code sends, surfacing a clear message when an account has no Copilot Auto available. Passes Codex requests to non-OpenAI providers through the vendor's Responses API exactly as Codex sends them, preventing invalid codex request rejections, and routes around a `gpt-6.x` refusal (including `bio_policy` safety-buffering failures) by trying other accounts for the same model before moving to the next account.

— Three fixes described without any exposed config. v0.1.498

30 **Model preference key validation and rename consistency**

IMPROVED

63

Adds `settings.checkModelKey` to validate per-model preference keys in the format `<provider id>/<model id>` or `<provider id>/*` at write time, rejecting unparseable keys rather than silently ignoring them. Adds a `renameModelPrefs` walk over all per-model maps in a loop so every per-model map (cost, size, upstream name) updates when a provider is renamed, not just the first matched.

— Internal setting names given, no user command. v0.1.496

THINNER COVERAGE BELOW

31 **'All' chip for bulk Library assignment** NEW 57

Adds an `All` chip at the start of each Library row so a single click assigns an MCP server or skill to every eligible agent, or removes it from all of them when all already have it.

— UI action described, no config surface named. v0.1.510

32 **Library-skill conflict resolution actions** NEW 57

Adds `UseLibrarySkill` and `KeepAgentSkill` resolution actions for settling library-skill conflicts when an agent already has a skill of its own by the same name.

— Names two actions, no config path shown. v0.1.498

33 **Codex compaction request-shape fix** IMPROVED 57

Codex compaction for a magpie model now requests its summary the same way Codex requests its own: streamed, with `tool_choice` set to `"auto"` and `parallel_tool_calls` set to `false`, so upstreams that validate Codex request shape no longer reject it.

— Field values named, no user action available. v0.1.470

34 **DeepSeek Harness low reasoning effort** IMPROVED 55

Extends DeepSeek Harness (`dsh`) thinking-effort to include `low` , giving the full set: `off` , `low` , `high` , `max` .

— Names effort levels, minimal mechanism detail. v0.1.455

35 **Qoder free-model labelling** IMPROVED 52

Displays Qoder models correctly as free when the model listing carries `is_free` / `isFree` true or a `price_factor` / `priceFactor` of 0, so practitioners can tell at a glance which Qoder models consume no credits.

— Fields named, but purely a display fix. v0.1.487

36 **macOS session terminal picker** NEW 52

Adds macOS Settings → Preferences → Session terminal picker, which selects which installed terminal app opens sessions from the terminal button in Usage → Sessions.

— UI path given, no further mechanism. product docs

37 **Windows title bar theming** IMPROVED 47

On Windows, the title bar is now painted to match the page's own background colour, following the active light or dark theme, so the bar and page read as a single surface rather than a grey strip above it.

— Visual change only, no config surface. v0.1.452

38 **Claude subscription image forwarding fix** IMPROVED 42

Forwards tool-returned images (PNG, browser screenshots) to the model on a Claude subscription, correcting a path where they previously arrived as text only.

— Bugfix described, no surface to act on. v0.1.490

39 **Responses client include-field propagation fix** IMPROVED

38

Propagates a Responses client's `include` field through to the Responses provider so the rebuilt request no longer drops it.

— Names the field but gives no user action. v0.1.490

40 **GLM Coding Plans wired to Codex Responses output** IMPROVED

35

Wires Codex Responses API output into GLM Coding Plans so the plan consumes model responses as-is.

— One-line description, no mechanism detail given. v0.1.490

41 **Routing list and model-effort display refinements** IMPROVED

35

Shows Codex app model effort figures sourced from magpie's own models, and adds DeepSeek chat GPT-call detail to the Routing list.

— Two bare UI mentions, no named surface. v0.1.490

└──▷ ALSO FROM THESE RELEASES

Override the billed price for a relay provider reselling GPT-5.5 at a discount so cost reports are accurate.

```
$ magpie model price relay-a/gpt-5.5 0.12,0.60,0.01,0.15
```

v0.1.515 CODE ↗ LATEST STABLE

Kimi Code (2.x) now receives per-model thinking levels via `support_efforts` and `default_effort` in its config.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.515 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.515
```

└──▷ USE IT

After upgrading, inspect the generated Kimi Code config to confirm thinking-level fields are populated for your models.

```toml
[models."<provider>/<model-id>"]
capabilities = ["thinking"]
support_efforts = ["low", "high", "max"]
default_effort = "high"
```

v0.1.514 CODE ↗ STABLE

Claude usage-limit resets surface on account cards; plugin accounts now expose full allowance data via `auth.usage`

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.514 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.514
```

Exposes Claude subscription usage-limit resets (queried via `/api/oauth/usage?cedar_ember=1&skip_spend=1`) on each Claude account card, usable manually or automatically when the week's quota is exhausted.

Plugin accounts now report allowance through `auth.usage`, surfacing `plan`, `user`, `until`, `renew`, `balance`, and `reset` fields — matching built-in account display parity.

Caps `maxTokens` sent to omp, Pi, Cline, and dsh agents at the model's `contextWindow` when the model's listed output limit exceeds its context window, preventing agents from requesting replies longer than the entire context.

v0.1.513 CODE ↗ STABLE

magpie v0.1.513 routes each model through its provider's native API so Claude thinking signatures survive tool turns.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.513 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.513
```

Adds `api` and `baseUrl` fields to `ompModel` in `models.yml` so each model is routed to its provider's native API (Anthropic Messages at `/v1/messages`, OpenAI Responses at `baseUrl/responses`, or Chat) instead of being written as `openai-completions`.

Routes Anthropic-served models through the gateway's `/v1/messages` endpoint, preserving Claude extended thinking signatures and `context_management`, `thinking.display`, and `output_config.effort` fields across tool turns.

v0.1.511 CODE ↗ STABLE

Magpie now appends `[1m]` to every 1M-context Claude model name it writes, preventing Claude Code from compacting prematurely.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.511 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.511
```

Automatically appends `[1m]` to any Claude model reference (in `model`, `env.ANTHROPIC_MODEL`, `env.ANTHROPIC_SMALL_FAST_MODEL`, and tier fields of `~/.claude/settings.json`) whose context window is 1 million tokens or more, so Claude Code 2.1.284 and earlier treat it as a 1M-context model instead of defaulting to 200K and triggering repeated autocompaction.

v0.1.510 CODE ↗ STABLE

Magpie v0.1.510 adds an `All` chip in the Library to assign an MCP server or skill to every eligible agent in one click.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.510 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.510
```

v0.1.509 CODE ↗ STABLE

Adds `X-Magpie-Account` header to pin gateway requests to a specific account within a subscription.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.509 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.509
```

└──▷ **TRY IT**

Pin a probe or test request to one specific account (by email or account ID) in a subscription that holds several, so you can isolate behavior per account.

```
$ curl -X POST https://<magpie-host>/v1/responses \
  -H 'X-Magpie-Account: me@example.com' \
  -H 'Content-Type: application/json' \
  -d '{"model":"gpt-5.5","input":"Hello"}'
```

v0.1.507 CODE ↗ STABLE

Adds opt-in auto-spend of Codex Bank Reset credits when a weekly quota is exhausted, persisted in `codex-autoreset.json`.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.507 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.507
```

Adds `CodexAutoReset` setting (stored in `settings.CodexAutoReset`) to allow a named Codex account to automatically spend one of its Bank Reset credits when its weekly window is used up, off by default.

Persists auto-reset state to `codex-autoreset.json` so a process restart does not spend an additional reset.

v0.1.504 CODE ↗ STABLE

magpie v0.1.504 adds per-provider/model price overrides with `magpie model price` and `magpie model prices`.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.504 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.504
```

└──▷ **TRY IT**

Set accurate per-token pricing for a relay provider that resells GPT models at a different rate than models.dev lists.

```
$ magpie model price relay-a/gpt-5.5 0.12,0.60,0.01,0.15
```

Apply a single flat price to every model under a provider with a wildcard, covering unlisted or custom models in bulk.

```
$ magpie model price relay-a/* 0.12,0.60,0.01,0.15
```

Audit all custom price overrides you have set across providers before a billing review.

```
$ magpie model prices
```

Adds `magpie model price <provider/model> <in>,<out>,<cache-read>,<cache-write>` to set a custom USD-per-million-token cost for any model or provider wildcard (e.g. `relay-a/*`), overriding models.dev list prices for resellers and unlisted models.

Adds `magpie model price <provider/model> --reset` to remove a user-set price and fall back to the lookup chain (per-model → `<provider>/*` → provider catalogue → models.dev).

Adds `magpie model prices` to list every model or provider pattern the user has explicitly priced.

v0.1.502 CODE ↗ STABLE

magpie v0.1.502 adds an OpenChamber agent with its own model-priority chain and dedicated config paths.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.502 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.502
```

Adds `openChamber` agent with its own row, reading settings from `$OPENCHAMBER_DATA_DIR` (defaulting to `~/.config/openchamber`) via

`preferences.json` and `settings.json`, and respecting `$OPENCODE_CONFIG_DIR` for the bundled OpenCode instance.

Applies a per-session model-priority chain for OpenChamber: project `defaultModel` in `settings.json` projects[], then profile `defaultModel` in `preferences.json`, then the agent's own model, then OpenCode's model — so OpenChamber's own default wins over the one magpie sets.

Supports `smallModelOverride` (when `smallModelUseDefault` is false) as the small-model selection for OpenChamber titles and commit messages, falling back to OpenCode's `small_model`.

v0.1.501 CODE ↗ STABLE

Codex signed in with an API key now uses magpie as its provider, with magpie's full model catalog available in the picker.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.501 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.501
```

When a request targets one of Codex's own models (e.g. `gpt-5.5`) and OpenAI is unreachable, the gateway now returns a 502 with a human-readable message explaining that OpenAI can't be reached, identifying the model as a Codex-native model, and prompting the user to pick one of magpie's models instead of surfacing a raw dial error.

v0.1.500 CODE ↗ STABLE

magpie v0.1.500 lets plugin providers hold multiple signed-in accounts, mirroring built-in subscription behaviour.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.500 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.500
```

Enables plugin providers to hold multiple signed-in accounts, stored in `plugin-auth.json` under the provider's id for the first account and `id#<slot>` for subsequent ones, matching how built-in subscriptions manage accounts.

Adds an `Accounts` field to the `Provider` struct (in `internal/plugin/providers.go`) exposing all signed-in accounts per plugin provider, each with `Key`, `Type`, `AccountID`, and `Hint` sub-fields.

Account scope is isolated per request via `AsyncLocalStorage` (added to `internal/plugin/host.js`) so a plugin's `getAuth`, `client.auth.set`, and loader always operate in the context of the specific account making the request.

v0.1.499 CODE ↗ STABLE

Menu bar now shows multiple subscriptions' usage side by side instead of just one

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.499 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.499
```

Extends Settings → Usage in the menu bar to display any number of ticked subscriptions side by side, each rendered as its provider logo with two quota windows stacked (e.g. `42%` over `18%`) on macOS, or as pipe-separated text (`42% · 18% | 10% · 5%`) on other platforms.

v0.1.498 CODE ↗ STABLE

magpie v0.1.498 adds manual group routing, smarter library-skill conflict resolution, and improved Copilot/gateway account fallback.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.498 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.498
```

Adds `UseLibrarySkill` and `KeepAgentSkill` resolution actions for settling library-skill conflicts when an agent already has a skill of its own by the same name.

Adds `Manual` routing mode for groups, letting each request go to a hand-picked model selected on the group's card.

Enables the Copilot backend to auto-negotiate the API version VS Code sends, and surfaces a clear message when an account has no Copilot Auto available.

Passes Codex requests to non-OpenAI providers through the vendor's Responses API exactly as Codex sends them, preventing `invalid codex request` rejections.

Routes around a `gpt-6.x` refusal (including `bio_policy` safety-buffering failures) by trying other accounts for the same model before moving to the next account.

v0.1.496 CODE ↗ STABLE

magpie v0.1.496 adds MCP server support for Hermes, omp, Kimi Code, Devin, Grok Build, Cline, and Command Code agents, plus per-model preference infrastructure.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.496 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.496
```

Adds MCP server integration for the Hermes agent via `mcp_servers` in its `config.yaml`, supporting streamable HTTP and SSE transports.

Adds MCP server integration for omp and Qoder agents via their `mcp.json`, with `type` field accepting `http` or `sse` for remote servers.

Adds MCP server integration for Kimi Code (both Python kimi-cli and new versions) via `mcp.json`, using `transport` field with `http` or `sse`.

Adds MCP server integration for Devin via `mcp_config.json`, writing a `transport` field (`stdio`, `http`, or `sse`) on every entry.

Adds MCP server integration for Grok Build via `[mcp_servers.<name>]` tables in `config.toml` (no SSE support).

+4 more

v0.1.490 CODE ↗ STABLE

magpie v0.1.490 adds `PI_CODING_AGENT_DIR` support, fixes Claude subscription image forwarding, and wires Codex responses into GLM Coding Plans.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.490 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.490
```

└──▷ **TRY IT**

Relocate Pi's agent config to a custom dotfiles directory so settings.json and models.json live outside the default ~/.pi/agent path.

```
$ export PI_CODING_AGENT_DIR=~/dotfiles/pi
```

Adds `PI_CODING_AGENT_DIR` environment variable to relocate Pi's agent folder (storing `settings.json` and `models.json`), with `~` `-expansion` support; relative paths and foreign-machine variables are ignored.

Forwards tool-returned images (PNG, browser screenshots) to the model on a Claude subscription, correcting a path where they previously arrived as text only.

Wires Codex Responses API output into GLM Coding Plans so the plan consumes model responses as-is.

Exposes Pi's local sessions under Usage › Sessions in the GUI.

Shows Codex app model effort figures sourced from magpie's own models.

+2 more

v0.1.489 CODE ↗ STABLE

magpie v0.1.489 adds an in-app Plugins tab and market backed by a `/api/plugins/market` endpoint for discovering and installing provider plugins.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.489 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.489
```

└──▷ **TRY IT**

Query the plugin market API directly to see what provider plugins are available, including community listings and live npm metadata.

```
$ curl http://localhost:<port>/api/plugins/market
```

Force magpie to use only the built-in bundled plugin registry without reaching out to GitHub — useful in air-gapped environments.

```
$ MAGPIE_PLUGIN_MARKET=off magpie
```

Point magpie at a self-hosted plugin registry instead of the default community repo.

```
$
MAGPIE_PLUGIN_MARKET=https://registry.internal.example.com/plugins/registry.json magpie
```

Adds a `GET /api/plugins/market` API endpoint that returns the community plugin registry (with live npm metadata), powering plugin discovery and search.

Adds a `MAGPIE_PLUGIN_MARKET` environment variable to override the default registry URL (`https://raw.githubusercontent.com/magpie-community/plugins/main/registry.json`); set it to `off` to force use of the built-in bundled copy.

Adds a Plugins tab to the GUI with a Discover view (community-suggested plugins with npm metadata) and an Installed view for managing, updating, enabling, or removing installed provider plugins.

v0.1.487 CODE ↗ STABLE

magpie v0.1.487 adds removal of signed-out Claude accounts from the Providers UI and correct free-model labelling for Qoder.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.487 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.487
```

Adds a `Remove` button in the Providers UI for Claude accounts that are saved in magpie but signed out (e.g. banned accounts), letting practitioners purge magpie's stored credentials without touching Claude Code's own files or keychain.

Displays Qoder models correctly as free when the model listing carries `is_free` / `isFree` true or a `price_factor` / `priceFactor` of 0, so practitioners can tell at a glance which Qoder models consume no credits.

v0.1.477 CODE ↗ STABLE

Routing groups can now explicitly name their own reasoning levels, preventing low-effort members from capping the whole group.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.477 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.477
```

Adds a `Levels` field on routing groups so a group can explicitly declare which reasoning levels (`none` , `low` , `medium` , `high` , `xhigh` , `max`) it advertises, overriding the intersection of what its members individually support.

v0.1.474 CODE ↗ STABLE

Magpie now syncs MCP servers to `Claude-3p/claude_desktop_config.json` and re-checks for updates immediately when a proxy is configured.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.474 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.474
```

Writes library MCP servers to `Claude-3p/claude_desktop_config.json` so Claude Desktop's 3p gateway mode picks them up alongside the standard `Claude/claude_desktop_config.json` .

Update checks now route through magpie's configured proxy, and a failed direct update check is retried immediately when a proxy is saved on the Settings page rather than waiting up to six hours.

v0.1.470 CODE ↗ STABLE

Magpie v0.1.470 wires in Kimi Code 2.x via ~/.kimi-code and fixes Codex compaction to stream correctly with tool_choice.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.470 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.470
```

Adds Kimi Code (2.x) support, reading its settings from `~/.kimi-code/config.toml` — distinct from the legacy `~/.kimi/config.toml` — so models selected in magpie are correctly routed to Kimi Code's `/model` endpoint.

Codex compaction for a magpie model now requests its summary the same way Codex requests its own: streamed, with `tool_choice` set to `"auto"` and `parallel_tool_calls` set to `false`, so upstreams that validate Codex request shape no longer reject it.

v0.1.469 CODE ↗ STABLE

magpie v0.1.469 extends skills delivery to 13 more agents via `~/.agents/skills` and skips agents whose home path is a file

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.469 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.469
```

Adds skills support for 13 more agents — including `kimi`, `goose`, `cindy`, `commandcode`, `devin`, `hermes`, `droid`, `cline`, `qoder`, `grok`, `workbuddy`, `fx`, and `devin` — routing each to its own skills folder or the shared `~/.agents/skills` directory.

Agents whose home path (`~/.dsh`, for example) is occupied by a file are now silently excluded from the library target list and skipped during writes rather than returning an 'open … not a directory' error.

v0.1.466 CODE ↗ STABLE

Agent model lists now show each model's maker logo, with new `logo` field in `agentModelJSON` and icons for Mimo/Xiaomi and StepFun model families.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.466 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.466
```

Adds a `logo` field (JSON: `logo`) to `agentModelJSON` so each model row in an agent's model list carries its maker's logo, resolved by model family or provider.

Exports ModelIcon(model string) in `internal/agent/icons.go` to look up a model's maker icon by model ID or name alone, returning `""` when the family is unrecognised.

Adds maker-logo recognition for `mimo` / `xiaomi` (icon: `mimocode`) and `step` (icon: `stepfun-color`) model families in the icon family table.

v0.1.463 CODE ↗ STABLE

Magpie v0.1.463 adds a Hide all button to the agent model list for bulk-hiding models in one click.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.463 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.463
```

v0.1.460 CODE ↗ STABLE

Magpie v0.1.460 adds per-agent model filtering, letting users hide models from individual agents' pickers via `settings.HiddenModels`.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.460 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.460
```

v0.1.459 CODE ↗ STABLE

Adds Antigravity (`agy`) as a Library target, wiring rules, MCP servers, and skills through `~/.gemini/config`

└──▷ GET THIS VERSION

```
$ git clone --branch v0.1.459 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.459
```

Adds `agy` as a named Library target, reading rules from `~/.gemini/config/GEMINI.md`, MCP servers from `~/.gemini/config/mcp_config.json`, and skills from `~/.gemini/config/skills`.

Adds `fmtAntigravity` MCP format support, encoding remote servers with `serverUrl` (plus `type: sse` for SSE transports) and decoding both `serverUrl` and fallback `url` fields from Antigravity's `mcp_config.json`.

v0.1.455 CODE ↗ STABLE

magpie v0.1.455 adds provider duplication, account pausing, per-key Balance URLs, and DeepSeek `low` reasoning effort.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.455 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.455
```

Adds `{key}` placeholder support in the Balance URL so each API key is named individually when checking balances.

Adds provider duplication via `AddCopy`, cloning a provider's keys, balance token, and routing alongside the original.

Adds the ability to pause a signed-in account on a provider without removing it.

Extends DeepSeek Harness (`dsh`) thinking-effort to include `low`, giving the full set: `off`, `low`, `high`, `max`.

v0.1.454 CODE ↗ STABLE

Adds OmO (oh-my-openagent / omo-ai) as a supported agent, wiring its `~/.omo/agent` config the same way magpie wires Pi.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.454 https://github.com/yetone/magpie.git
# already have the repo? check out this version:
$ git checkout v0.1.454
```

Adds support for the `OMO_CODING_AGENT_DIR` and `SENPI_CODING_AGENT_DIR` environment variables to relocate OmO's agent folder away from the default `~/.omo/agent`.

Adds OmO (`omo-ai` / `oh-my-openagent`) as a new managed agent: magpie sets `defaultProvider`, `defaultModel`, and `defaultThinkingLevel` in `~/.omo/agent/settings.json` and registers itself as the `magpie` provider in `~/.omo/agent/models.json`, mirroring Pi integration without touching `~/.pi`.

Supports OmO's MCP server config at `~/.omo/agent/mcp.json` (Pi-native format), agent instructions at `~/.omo/agent/AGENTS.md`, and skills at `~/.omo/agent/skills`.

v0.1.452 CODE ↗ STABLE

> Magpie v0.1.452 paints the Windows title bar to match the app's light/dark theme colour.
>
> └──▷ **GET THIS VERSION**
>
> ```
> $ git clone --branch v0.1.452 https://github.com/yetone/magpie.git
> # already have the repo? check out this version:
> $ git checkout v0.1.452
> ```

WAS THIS USEFUL? 👍 👎

Cline

Feedback ⓘ desktop-v0.0.39 NOTES ↗ CODE ↗ RANK

Autonomous coding agent as an SDK, IDE extension, or CLI assistant.

> Cline shipped new Anthropic refusal fallback and content-filter handling across the extension and SDK, added Bee and Pareto Inference as LLM providers, refreshed default models for over a dozen providers (while dropping the Claude Opus 4 variants), and extended codebase search, hub diagnostics, and Linux desktop packaging.

└──▷ **WHAT SHIPPED · 13 FEATURES** most completely described first ? what's the number?

01 **Linux desktop packaging and rendering fixes** NEW 82

Adds a `WEBKIT_DISABLE_DMABUF_RENDERER=0` environment variable override to keep WebKitGTK's DMA-BUF renderer enabled on Linux hybrid-GPU systems, while automatically disabling that renderer when the proprietary NVIDIA driver is detected (presence of `/sys/module/nvidia`) to prevent blank app windows. Linux `.deb` and `.rpm` packages now ship AppStream metainfo files so GNOME Software and KDE Discover can show Cline's icon, description, and installed status, and new `arm64` `.deb` / `.rpm` packages are available for Linux ARM64 with auto-update support matching the x64 packages.

— Names exact env var, sysfs path, and package formats but no test recipe.

desktop-v0.0.39 desktop-v0.0.38

02 **Hub client version and PID diagnostics** NEW 80

The Hub web app now displays each connected client's version and PID for diagnostics; the SDK's `NodeHubClient` registers `{ version, pid }` metadata (version defaults to the `@cline/core` version, overridable with `clientVersion`), included in `hub.client.registered` events and visible for connected clients and join/leave events.

— Names exact client fields, event, and override option. sdk/sdk/v0.0.88 cli-v3.0.66

03 **Model catalog refresh and Claude Opus 4 removal** BREAKING 78

Refreshes the model catalog: `GPT-6.1 Sol` becomes the default for OpenAI, OpenRouter, GitHub Copilot, Cortecs, Eden AI, Kilo Gateway, both LLM Gateway providers, NanoGPT, OpenCode Zen, and Requesty; Vercel AI Gateway moves to Ling 3.1 Flash; Tempr Gateway to MiMo V2.6 Flash; CrossModel and Ofox to Claude Sonnet 5.5; Pioneer to GLiNER 2.5 Decide; Scaleway to Qwen 3.8 27B; and the Cline recommended list gains Claude Sonnet 5.5 and Claude Opus 5.5. Users who have not pinned a model

will automatically switch to these new defaults on upgrade, and the SDK separately drops the Claude Opus 4 variants from the model catalog.

— Lists every affected provider and new default model by name. `cli-v3.0.66` `v4.1.22` `sdk/sdk/v0.0.88`

04 **Content-filter and refusal handling across extension and SDK**

IMPROVED 75

Enables Anthropic server-side refusal fallback so that, on OpenRouter and Cline, Anthropic model requests can fail over to another upstream provider instead of failing outright, and surfaces an explicit content-filter message with a rephrase suggestion, replacing the generic `Model returned empty response` error. The SDK adds `content-filter` as a new `AgentModelFinishReason`, surfaced in both the AI SDK and ApiHandler adapters, so the `sdk.error` run-failed event carries the turn's finish reason instead of reporting an empty response.

— Names exact SDK event and enum but no usage example. `sdk/sdk/v0.0.88` `v4.1.22`

05 **New LLM providers: Bee and Pareto Inference** NEW 75

Adds Bee provider (by HEOSSI) as a new LLM backend option, and adds Pareto Inference provider configurable via `PARETO_API_KEY`.

Connect to the Pareto Inference provider by setting its API key before running a Cline session.

```
$ export PARETO_API_KEY=<your-pareto-key>
cline "Refactor the auth module to use async/await"
```

— Runnable example sets `PARETO_API_KEY` directly. `v4.1.22`

06 **PHP support in codebase search** NEW 75

Extends `search_codebase` to index PHP sources and templates (`php`, `phtml`, `inc`, `twig`) and excludes Composer's `vendor` directory the same way `node_modules` is excluded.

— Names exact file extensions and excluded directory. `cli-v3.0.66` `desktop-v0.0.38`

07 **Yolo mode skips plan/act system prompt injection** IMPROVED 70

Yolo mode (`-y`) no longer injects plan/act mode instructions into the system prompt, keeping it lean for scripted use.

Run a fully autonomous session without plan/act instructions polluting the system prompt, keeping it lean for scripted CI use.



— Comes with a runnable CLI example. cli-v3.0.66

08 **UTF-8 BOM support for scheduled task and cron specs** IMPROVED

60

Supports cron (`.cline/cron/*.md`) and task (`*.task.md`) spec files saved with a UTF-8 BOM, as produced by Windows Notepad, so they now parse correctly.

— Names exact file patterns affected by the fix. cli-v3.0.66 desktop-v0.0.38

THINNER COVERAGE BELOW

09 **Web search settings moved to Customize → Tools** IMPROVED

55

Moves web search settings from General to Customize → Tools, showing provider search support status and linking to provider settings when unsupported.

— Gives exact UI navigation path but no further mechanism. desktop-v0.0.38

10 **CLI binaries rebuilt with Bun 1.4.2** IMPROVED

55

CLI binaries rebuilt with Bun 1.4.2, enabling launch on macOS 27 and execution on x64 CPUs without AVX2.

— Names runtime version and platform support but no command. cli-v3.0.66

11 **Cloud agent session workspace persistence** IMPROVED

45

Cloud agent sessions now preserve their workspace — a suspended cloud session resumes with the same files, history, and reasoning settings.

— Describes behavior but no config or command to control it. desktop-v0.0.38

12 **Connectors marked Beta with filtered marketplace view** IMPROVED

40

Connectors are now labeled Beta, and opening the marketplace from Connectors shows only connectors.

— Short UI-label change, no further mechanism. desktop-v0.0.38

13 **Gateway models keep dedicated API protocol** IMPROVED

30

Gateway models now keep their own API protocol instead of falling back to the provider-wide default.

— Bare description, no named protocol or config. desktop-v0.0.38

└──▷ **BREAKING ON UPGRADE**

! Default models change for many providers on upgrade: if you use OpenAI, OpenRouter, GitHub Copilot, Cortecs, Eden AI, Kilo Gateway, either LLM Gateway provider, NanoGPT, OpenCode Zen, Requesty, Vercel AI Gateway, Tempr Gateway, CrossModel, Ofox, Pioneer, or Scaleway without pinning a model, your active default will switch to a new model automatically.

! Default model changes automatically for users of OpenAI, OpenRouter, GitHub Copilot, Cortecs, Eden AI, Kilo Gateway, both LLM Gateway providers, NanoGPT, OpenCode Zen, and Requesty who have not pinned a model — they will now receive GPT-6.1 Sol. Vercel AI Gateway moves to Ling 3.1 Flash, Tempr Gateway to MiMo V2.6 Flash, CrossModel and Ofox to Claude Sonnet 5.5, Pioneer to GLiNER 2.5 Decide, and Scaleway to Qwen 3.8 27B.

! The Claude Opus 4 variants are dropped from the model catalog.

desktop-v0.0.39 NOTES ↗ CODE ↗ LATEST STABLE

Cline desktop v0.0.39 adds Linux NVIDIA rendering fix with `WEBKIT_DISABLE_DMABUF_RENDERER` override and AppStream metadata for GNOME Software and KDE Discover.

└──▷ **GET THIS VERSION**

```
$ git clone --branch desktop-v0.0.39
https://github.com/cline/cline.git
# already have the repo? check out this version:
$ git checkout desktop-v0.0.39
```

Adds `WEBKIT_DISABLE_DMABUF_RENDERER=0` environment variable override to keep WebKitGTK's DMA-BUF renderer enabled on Linux hybrid-GPU systems where the NVIDIA driver is present but DMA-BUF should remain active.

Automatically disables WebKitGTK's DMA-BUF renderer on Linux when the proprietary NVIDIA driver is detected (presence of `/sys/module/nvidia`), preventing blank app windows on affected systems.

Linux `.deb` and `.rpm` packages now ship AppStream metainfo files, enabling GNOME Software and KDE Discover to display Cline's icon, description, and installed status.

v4.1.22 NOTES ↗ CODE ↗ STABLE

Adds Bee and Pareto Inference providers (via `PARETO_API_KEY`) plus Anthropic server-side failover and refreshed model defaults.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v4.1.22 https://github.com/cline/cline.git
# already have the repo? check out this version:
$ git checkout v4.1.22
```

└──▷ **TRY IT**

Connect to the Pareto Inference provider by setting its API key before running a Cline session.

```
$ export PARETO_API_KEY=<your-pareto-key>
cline "Refactor the auth module to use async/await"
```

Adds Bee provider (by HEOSSI) as a new LLM backend option.

Enables Anthropic server-side refusal fallback so that, on OpenRouter and Cline, Anthropic model requests can fail over to another upstream provider instead of failing outright.

Surfaces an explicit content-filter message with a rephrase suggestion when a response is blocked, replacing the generic `Model returned empty response` error.

Refreshes the model catalog: GPT-6.1 Sol becomes the default for OpenAI, OpenRouter, GitHub Copilot, Cortecs, Eden AI, Kilo Gateway, both LLM Gateway providers, NanoGPT, OpenCode Zen, and Requesty; Vercel AI Gateway moves to Ling 3.1 Flash; Tempr Gateway to MiMo V2.6 Flash; CrossModel and Ofox to Claude Sonnet 5.5; Pioneer to GLiNER 2.5 Decide; Scaleway to Qwen 3.8 27B; and the Cline recommended list gains Claude Sonnet 5.5 and Claude Opus 5.5.

└──▷ **BREAKING ON UPGRADE**

! Default models change for many providers on upgrade: if you use OpenAI, OpenRouter, GitHub Copilot, Cortecs, Eden AI, Kilo Gateway, either LLM Gateway provider, NanoGPT, OpenCode Zen, Requesty, Vercel AI Gateway, Tempr Gateway, CrossModel, Ofox, Pioneer, or Scaleway without pinning a model, your active default will switch to a new model automatically.

desktop-v0.0.38 NOTES ↗ CODE ↗ STABLE

Cline desktop v0.0.38 adds Linux ARM64 packages, Anthropic refusal fallback, two new providers, and refreshed model defaults.

└──▷ **GET THIS VERSION**

```
$ git clone --branch desktop-v0.0.38
https://github.com/cline/cline.git
# already have the repo? check out this version:
$ git checkout desktop-v0.0.38
```

Ships `arm64` `.deb` and `.rpm` packages for Linux on ARM64, with auto-update support matching the x64 packages.

Moves web search settings from `General` to `Customize → Tools`, showing provider search support status and linking to provider settings when unsupported.

Cloud agent sessions now preserve their workspace — a suspended cloud session resumes with the same files, history, and reasoning settings.

Gateway models now keep their own API protocol instead of falling back to the provider-wide default.

Supports scheduled task specs saved with a UTF-8 BOM (as Windows Notepad produces), and extends codebase search to include PHP projects (skipping Composer's `vendor` folder).

+1 more

`cli-v3.0.66` `NOTES ↗` `CODE ↗` `STABLE`

Cline: Extends `search_codebase` to index PHP sources and templates (`php` , `phtml` , `inc` , `twig`), and excludes Composer's `vendor` directory the same way `node_modules` is excluded.

└──▷ **GET THIS VERSION**

```
$ git clone --branch cli-v3.0.66 https://github.com/cline/cline.git
# already have the repo? check out this version:
$ git checkout cli-v3.0.66
```

└──▷ **TRY IT**

Run a fully autonomous session without plan/act instructions polluting the system prompt, keeping it lean for scripted CI use.

```
$ cline -y "Run the test suite and fix any failures"
```

Hub web app now displays each connected client's version and PID for diagnostics.

Yolo mode (`-y`) no longer injects plan/act mode instructions into the system

prompt.

Cron (`.cline/cron/*.md`) and task (`*.task.md`) spec files saved with a UTF-8 BOM now parse correctly.

CLI binaries rebuilt with Bun 1.4.2, enabling launch on macOS 27 and execution on x64 CPUs without AVX2.

└──▷ **BREAKING ON UPGRADE**

! Default model changes automatically for users of OpenAI, OpenRouter, GitHub Copilot, Cortecs, Eden AI, Kilo Gateway, both LLM Gateway providers, NanoGPT, OpenCode Zen, and Requesty who have not pinned a model — they will now receive GPT-6.1 Sol. Vercel AI Gateway moves to Ling 3.1 Flash, Tempr Gateway to MiMo V2.6 Flash, CrossModel and Ofox to Claude Sonnet 5.5, Pioneer to GLiNER 2.5 Decide, and Scaleway to Qwen 3.8 27B.

sdk/sdk/v0.0.88 NOTES ↗ CODE ↗ STABLE

Cline SDK v0.0.88 adds provider fallback routing, a `content-filter` finish reason, and `NodeHubClient` version/PID metadata.

└──▷ **GET THIS VERSION**

```
$ git clone --branch sdk/sdk/v0.0.88
https://github.com/cline/cline.git
# already have the repo? check out this version:
$ git checkout sdk/sdk/v0.0.88
```

Adds `content-filter` as a new `AgentModelFinishReason`, surfaced in both AI SDK and ApiHandler adapters, so content-filtered turns fail with an actionable message and the `sdk.error` run-failed event carries the turn's finish reason instead of reporting an empty response.

Adds `{ version, pid }` metadata registration to `NodeHubClient` (version defaults to the `@cline/core` version, overridable with `clientVersion`), included in `hub.client.registered` events and visible in the Cline Hub web app for connected clients and join/leave events.

└──▷ **BREAKING ON UPGRADE**

! The Claude Opus 4 variants are dropped from the model catalog.

WAS THIS USEFUL? 👍 👎

Letta Code

Feedback ⓘ v0.34.0 NOTES ↗ RANK

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

> Letta Code shipped async user-question handling, persistent client preferences, a storage-backed OAuth authorizer for MCP, and restricted the Memory Palace skill to Cloud.

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

01 **Storage-backed OAuth authorizer for MCP** NEW 38

Adds a storage-backed OAuth authorizer for MCP (Model Context Protocol) integration.

– Names the integration and mechanism but no config or endpoint detail v0.33.7

02 **Asynchronous user questions** IMPROVED 31

Agents no longer block waiting for human replies to user questions; the interaction is now asynchronous so the agent can continue working.

– Describes behavior change but no mechanism or config detail v0.34.0

03 **Persistent conversation client preferences** IMPROVED 31

Conversation client preferences now persist across sessions, so users do not need to reconfigure them on each restart.

– Clear behavior described, no specific keys or commands named v0.34.0

04 **Memory Palace restricted to Cloud** BREAKING 25

The Memory Palace skill is now restricted to Cloud-only deployments.

– Names the feature and restriction but no migration guidance v0.34.0

v0.34.0 NOTES ↗ LATEST STABLE

Letta Code v0.34.0 makes user questions asynchronous, persists conversation client preferences, and restricts Memory Palace to Cloud.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.34.0 https://github.com/letta-ai/letta-code.git
# already have the repo? check out this version:
$ git checkout v0.34.0
```

Makes user questions asynchronous so agents no longer block waiting for human replies.

Persists conversation client preferences across sessions so users do not need to reconfigure them on each restart.

Restricts the Memory Palace skill to Cloud-only deployments.

v0.33.7 NOTES ↗ STABLE

Letta Code v0.33.7 publishes a storage-backed OAuth authorizer for MCP.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.33.7 https://github.com/letta-ai/letta-code.git
# already have the repo? check out this version:
$ git checkout v0.33.7
```

Adds a storage-backed OAuth authorizer for MCP (Model Context Protocol) integration.

WAS THIS USEFUL? 👍 👎

Google gemini-cli

Feedback ⓘ v0.62.0 NOTES ↗ RANK

An open-source command-line AI agent for coding, task automation, and tool use with Gemini models.

> Gemini CLI added support for two new Gemini model variants, expanding the model options available to users.

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

01 Support for Gemini 3.8 Flash and 3.5 Flash Lite models NEW

60

Adds support for the `gemini-3.8-flash` and `gemini-3.5-flash-lite` models, allowing them to be selected for CLI sessions, e.g. via `gemini -m gemini-3.8-flash -p "..."`.

Run a prompt against the new Gemini 3.8 Flash model for faster, higher-capability responses.

```
$ gemini -m gemini-3.8-flash -p "Review this code for security issues"
```

— Names exact model identifiers and a runnable command. v0.62.0

v0.62.0 NOTES ↗ LATEST STABLE

Gemini CLI v0.62.0 adds Gemini 3.8 Flash and Gemini 3.5 Flash Lite model support.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.62.0 https://github.com/google-gemini/gemini-cli.git
# already have the repo? check out this version:
$ git checkout v0.62.0
```

└──▷ **TRY IT**

Run a prompt against the new Gemini 3.8 Flash model for faster, higher-capability responses.

```
$ gemini -m gemini-3.8-flash -p "Review this code for security issues"
```

> Adds support for `gemini-3.8-flash` and `gemini-3.5-flash-lite` models.

WAS THIS USEFUL? 👍 👎

OpenAI Codex CLI

Feedback ⓘ rust-v0.161.0-alpha.4 NOTES ↗

CODE ↗ RANK

OpenAI Codex CLI runs an agent in the terminal that reads, changes, and tests code in local repositories.

> Codex CLI's latest releases add local audio device selection for TUI voice conversations, expand MCP authentication with a named login command and enterprise sign-in, set GPT-6.1 Sol as the default model across bundled and Bedrock catalogs, and round out several smaller multi-agent and TUI additions.

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

01 Local audio device selection in TUI voice settings NEW

86

Codex CLI adds `audio.microphone` and `audio.speaker` config keys to persist local input/output device selections for voice conversations in the TUI, along with input and output device pickers in the TUI voice settings UI that include back navigation and a System default option.

Persist a specific microphone and speaker for all voice conversations so Codex does not default to the system audio device.

toml

```
audio.microphone = "Blue Yeti"
audio.speaker = "External Headphones"
```

— Names exact config keys and includes a runnable config example. rust-v0.161.0-alpha.4

02 MCP authentication and authorization NEW

78

Codex CLI adds MCP authorization server discovery and validation before ID-JAG exchange, a `/mcp login <name>` command in the TUI to authenticate MCP servers by name, and enterprise MCP sign-in with account-scoped OAuth grant cleanup on logout.

Authenticate a named MCP server from within the Codex TUI interactive session.

```
$ /mcp login <name>
```

— Names the login command and OAuth cleanup mechanism with a runnable example.

rust-v0.161.0-alpha.2 rust-v0.161.0-alpha.4

03 Analytics default flag for PID backend app server NEW

60

Codex CLI passes `--analytics-default-enabled` when the PID backend launches an app server, while still honoring `analytics.enabled = false` in `config.toml`.

— Names flag and config key but gives no usage example. rust-v0.161.0-alpha.4

THINNER COVERAGE BELOW

04 **GPT-6.1 Sol as default model** IMPROVED 40

Codex CLI sets GPT-6.1 Sol as the default model in the bundled catalog and in the Amazon Bedrock Mantle and Bedrock Runtime catalogs.

— Names catalogs affected but no migration or config guidance. rust-v0.159.1

05 **Multi-agent session enhancements** NEW 27

Codex CLI adds an opt-in model catalog to multi-agent context and supports remote agent message boards in multi-agent sessions.

— Two thin additions to the multi-agent area with no mechanism explained.

rust-v0.161.0-alpha.4 rust-v0.161.0-alpha.2

06 **Fork shortcut in TUI command center** NEW 20

Adds a fork shortcut to the TUI command center.

— Bare name with no navigation or behavior detail. rust-v0.161.0-alpha.4

07 **Experimental thread prediction protocol types** NEW 13

Adds experimental thread prediction protocol types.

— Bare name only, no explanation of behavior or use. rust-v0.161.0-alpha.4

rust-v0.161.0-alpha.4 NOTES ↗ CODE ↗ PRE-RELEASE

Codex CLI adds local audio device selection in TUI voice settings with `audio.microphone` and `audio.speaker` config keys.

└──▷ GET THIS VERSION

```
$ git clone --branch rust-v0.161.0-alpha.4
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.161.0-alpha.4
```

└──▷ USE IT

Persist a specific microphone and speaker for all voice conversations so Codex does not default to the system audio device.

config.toml

```
audio.microphone = "Blue Yeti"
audio.speaker = "External Headphones"
```

Adds `audio.microphone` and `audio.speaker` config keys to persist local input/output device selections for voice conversations in the TUI.

Adds input and output device pickers to the TUI voice settings UI, with back navigation and a System default option.

Passes `--analytics-default-enabled` when the PID backend launches an app server, while still honoring `analytics.enabled = false` in `config.toml`.

Adds a fork shortcut to the TUI command center.

Adds an opt-in model catalog to multi-agent context.

+2 more

rust-v0.159.1 NOTES ↗ CODE ↗ LATEST STABLE

Codex CLI v0.159.1 adds GPT-6.1 Sol as the default model across the bundled, Amazon Bedrock Mantle, and Bedrock Runtime catalogs.

└──▷ GET THIS VERSION

```
$ git clone --branch rust-v0.159.1
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.159.1
```

rust-v0.161.0-alpha.2 NOTES ↗ CODE ↗ PRE-RELEASE

Adds `/mcp login <name>` to the TUI and enables enterprise MCP sign-in with account-scoped grant cleanup.

└──▷ GET THIS VERSION

```
$ git clone --branch rust-v0.161.0-alpha.2
https://github.com/openai/codex.git
# already have the repo? check out this version:
$ git checkout rust-v0.161.0-alpha.2
```

└──▷ TRY IT

Authenticate a named MCP server from within the Codex TUI interactive session.

```
$ /mcp login <name>
```

Adds `/mcp login <name>` command to the TUI for authenticating MCP servers by name.

Enables enterprise MCP sign-in and account-scoped OAuth grant cleanup on logout.

Supports remote agent message boards in multi-agent sessions.

WAS THIS USEFUL? 👍 👎

Anthropic Claude Code

Feedback ⓘ v2.1.285 NOTES ↗

CODE ↗ RANK

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

> Claude Code's v2.1.285 release adds a desktop-app launcher, new plugin configuration tooling, managed policy controls for providers and plugin mods, two environment variables, and a model-fallback improvement for Bedrock and Vertex AI.

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

01 **Inline MCP server config on plugin install** NEW 85

`claude plugin install --config` gains a `<server>.<key>=<value>` syntax so a bundled `.mcpb` MCP server's settings can be configured at install time without visiting `/plugin` → Configure.

Install a plugin and pre-configure its bundled MCP server in one step, skipping the post-install Configure prompt.

```
$ claude plugin install my-plugin --config myserver.api_key=sk-abc123 --config myserver.region=us-east-1
```

— Exact syntax plus a runnable install command v2.1.285

02 **Environment variables for WebFetch and timeouts** NEW 75

Adds `CLAUDE_CODE_DISABLE_WEB_FETCH` to disable the WebFetch tool entirely, and `CLAUDE_CODE_NONSTREAMING_TIMEOUT_RETRIES` to cap re-sends of a non-streaming fallback request that timed out.

Disable WebFetch in a CI environment where outbound HTTP from the agent is not permitted.

```
$ CLAUDE_CODE_DISABLE_WEB_FETCH=1 claude -p 'Review the codebase and summarize open TODOs'
```

— Both vars named verbatim with a runnable example; no default values given v2.1.285

03 **claude plugin configure subcommand** NEW 75

`claude plugin configure <plugin>` inspects a plugin's options and shows which are unset, and can write new values from stdin via `--values-stdin`.

— Exact subcommand and flag named, no worked example v2.1.285

04 **Managed policy settings for providers and plugin mods** NEW

75

Adds `allowedProviders` managed setting restricting which API providers a machine may use (Anthropic API, custom endpoint, Bedrock, Mantle, Vertex AI, Foundry, Claude Platform on AWS, or a Cloud gateway), and `allowManagedModsOnly` under `pluginConfigs` to allow org-managed mods while blocking user-installed mods from loading.

— Both settings and provider list named, no config file path or example v2.1.285

05 **`claude --desktop` flag** NEW

70

`claude --desktop` opens the Claude desktop app on the current directory, or on a specific session using `--continue` or `--resume <id>`.

— Exact flag and its session options given, no example command v2.1.285

THINNER COVERAGE BELOW

06 **Model fallback for Bedrock and Vertex AI** IMPROVED

45

Bedrock and Vertex AI sessions now fall back to an older available model of the same tier when an admin removes access to the default model, instead of failing.

— Behavior change described but no config surface or trigger detail v2.1.285

v2.1.285 NOTES ↗ CODE ↗ LATEST STABLE

Adds `claude --desktop`, `claude plugin configure`, `allowedProviders` policy, and two new env vars for WebFetch and timeout control.

└──▷ GET THIS VERSION

```
$ git clone --branch v2.1.285 https://github.com/anthropics/claude-code.git
# already have the repo? check out this version:
$ git checkout v2.1.285
```

└──▷ TRY IT

Disable WebFetch in a CI environment where outbound HTTP from the agent is not permitted.

```
$ CLAUDE_CODE_DISABLE_WEB_FETCH=1 claude -p 'Review the codebase and summarize open TODOs'
```

Install a plugin and pre-configure its bundled MCP server in one step, skipping the post-install Configure prompt.

```
$ claude plugin install my-plugin --config myserver.api_key=sk-abc123 --config myserver.region=us-east-1
```

Adds `CLAUDE_CODE_DISABLE_WEB_FETCH` environment variable to disable the WebFetch tool entirely.

Adds `CLAUDE_CODE_NONSTREAMING_TIMEOUT_RETRIES` environment variable to cap re-sends of a non-streaming fallback request that timed out.

Adds `claude --desktop` flag to open the Claude desktop app on the current directory, or on a specific session with `--continue` or `--resume <id>`.

Adds `claude plugin configure <plugin>` subcommand to inspect a plugin's options and which are unset, or write new values from stdin via `--values-stdin`.

Adds `<server>.<key>=<value>` syntax to `claude plugin install --config` so a bundled `.mcpb` MCP server's settings can be configured at install time without visiting `/plugin` → Configure.

+3 more

WAS THIS USEFUL? 👍 👎

Graphify Labs Graphify

Feedback ⓘ v0.9.72 NOTES ↗

RANK

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

> Graphify v0.9.72 adds automatic self-healing for stale skill files and extends its knowledge graph to understand Blade template inheritance.

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

01 **Auto-refresh of stale installed skills** NEW 95

After a package upgrade, Graphify now automatically refreshes stale installed skills, triggered on any non-install CLI command. It backs up local edits to `SKILL.md.bak`, preserves marker-bounded `CLAUDE.md` / `AGENTS.md` / `GEMINI.md` sections, and the behavior can be disabled by setting `GRAPHIFY_NO_AUTO_REFRESH=1`.

Suppress automatic skill refresh in environments where you manage `SKILL.md` manually, such as locked CI containers.

```
$ GRAPHIFY_NO_AUTO_REFRESH=1 graphify query "APIRouter"
```

— Names exact files, env var and trigger condition; includes runnable example v0.9.72

THINNER COVERAGE BELOW

02 **Blade template layout linking** NEW 45

Blade templates now link a view to the layout it `@extends`, adding cross-template dependency edges to the knowledge graph.

— Explains mechanism but no command or navigation path given v0.9.72

v0.9.72 NOTES ↗ LATEST STABLE

Graphify v0.9.72 adds auto-refresh of stale skills on any CLI command and Blade template layout linking.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.9.72 https://github.com/Graphify-Labs/graphify.git
# already have the repo? check out this version:
$ git checkout v0.9.72
```

└──▷ **TRY IT**

Suppress automatic skill refresh in environments where you manage `SKILL.md` manually, such as locked CI containers.

```
$ GRAPHIFY_NO_AUTO_REFRESH=1 graphify query "APIRouter"
```

Adds automatic refresh of stale installed skills after a package upgrade, triggered on any non-install CLI command; backs up local edits to `SKILL.md.bak`, preserves marker-bounded `CLAUDE.md` / `AGENTS.md` / `GEMINI.md` sections, and can be disabled with `GRAPHIFY_NO_AUTO_REFRESH=1`.

Blade templates now link a view to the layout it `@extends`, adding cross-template dependency edges to the knowledge graph.

WAS THIS USEFUL? 👍 👎

Earendil Works Pi

Feedback ⓘ v0.99.1 NOTES ↗ RANK

AI agent toolkit: unified LLM API, agent loop, TUI, coding agent CLI

> Pi v0.99.0 adds first-class MCP server support and a QuickJS-based codemode sandbox for model-written JavaScript, alongside virtual models, ChatGPT sign-in, and a new GPT-6.1 Sol default for OpenAI Codex, with several breaking changes to built-in extension naming, disabling behaviour, and defaults.

└──▷ **WHAT SHIPPED · 19 FEATURES** most completely described first ? what's the number?

01 **MCP server support** NEW 95

Pi now supports Model Context Protocol servers over stdio or streamable HTTP with OAuth. Configure servers globally or per project via `mcp.json`, register them programmatically with pi.registerMcpServer(), and manage them with `pi mcp add|remove|list|login|logout` or the `/mcp` slash command.

— Full transport, config file, CLI and slash command named v0.99.0

02 **Built-in extension management and naming** BREAKING 90

Adds a Built-in section in `pi config` to disable built-in extensions (`mcp` , `llama.cpp` , `codemode` , `tool-search`) globally or per project using `-builtin:<name>` in the `extensions` setting. `--no-extensions` now also disables built-in extensions including the llama.cpp provider (previously they stayed active); restore one explicitly with `-e builtin:<name>` (e.g. `pi -ne -e builtin:mcp`). Built-in extensions and tools are now named `builtin:<name>` (e.g. `builtin:mcp` , `builtin:read`) in errors, diagnostics, RPC source info, and bug reports, replacing the previous `<inline:name>` / `<builtin:name>` naming.

— Config section, flags and naming scheme all named v0.99.0

03 **codemode JavaScript sandbox extension** NEW 85

Adds a `codemode` built-in extension that runs model-written JavaScript in a QuickJS sandbox. Enable it with `--tools` or `defaultTools` (e.g. `"+codemode"`) and configure it with `codemode.mode` and `codemode.inlineBudget` in project config.

Persist codemode as a default tool and cap its inline budget in your project config.

```json
{
  "defaultTools": ["+codemode"],
  "codemode.mode": "auto",
  "codemode.inlineBudget": 4096
}
```

— Enable flag and both config keys named with example v0.99.0

04 **Extension tool API fields and nested tool calls** NEW

85

Extension tool definitions gain new fields: `exposure` (`direct`, `model-only`, `codemode`, `deferred`, or `hidden`), `namespace`, `annotations`, `outputSchema` with `structuredContent`, and `isError` results, plus a prepareLoadout() hook. Extensions can now call ctx.executeTool() to make nested tool calls, emitting events tagged with `parentToolCallId` and recording bounded `nestedCalls` on the calling tool's result.

— Names every new field and method with mechanism v0.99.0

05 **GPT-6.1 Sol model and Codex provider rename** BREAKING

80

Adds `gpt-6.1-sol` to the OpenAI, Azure OpenAI, and OpenAI Codex providers and makes it the new default OpenAI Codex model, replacing the previous default; existing workflows relying on the old default will now use GPT-6.1 Sol unless a model is explicitly specified. The inherited OpenAI Codex provider is also renamed to `OpenAI Codex (legacy)`, so any configuration referencing it by name may need updating.

— Model id, providers and rename named with before/after v0.99.1

06 **Larger codemode command output handling** IMPROVED

75

In codemode, `bash` and `powershell` structured results can now hold up to 1 MiB of output, up from a previous cap of 2000 lines or 50 KB, and results add `truncated` and `full_output_path` fields.

— Concrete before/after limits and new fields named v0.99.0

07 **Virtual models** NEW

75

Adds experimental virtual models: extensions register them with pi.registerVirtualModel() to route each request to a different physical model and thinking level; `/session` lists cost per physical model.

— Named API and UI surface, mechanism described v0.99.0

08 **ModelRuntime API extensions for images and classification** NEW 75

Adds generateImages() to `ModelRuntime` with runtime-resolved auth, plus getModelsOfType(), getModelOfType(), getAvailableOfType(), getAllModels(), and getAllAvailable(). Also adds classify() and classifier model accessors, including the built-in `jev-latest` model; classifier models are additionally available for every llama.cpp chat model via next-token label probabilities.

— Every new method and model named, thin usage detail v0.99.0

09 **defaultTools +/- shorthand** NEW 65

Adds `+name` and `-name` entries to the `defaultTools` setting to add or remove tools without repeating all defaults, e.g. `{"defaultTools": ["+codemode"]}`.

Enable codemode by default for every session without touching the full defaultTools list.

```json
{
  "defaultTools": ["+codemode"]
}
```

— Runnable config example with exact syntax v0.99.0

10 **ChatGPT sign-in for OpenAI provider** NEW 60

Adds `/login openai` to sign in with ChatGPT, letting users authenticate with a ChatGPT subscription against the OpenAI provider.

— Exact command named, minimal further mechanism v0.99.0

11 **Fullscreen mouse-wheel scroll setting** NEW 60

Adds `fullscreenWheelScrollLines` setting and a `/settings` entry for configuring fullscreen mouse-wheel scroll speed; defaults to `"auto"`, which accelerates fast spins.

— Config key, UI path and default value named v0.99.0

12 **RPC per-input disposition** IMPROVED 60

Adds per-input disposition to RPC `prompt`, `steer`, and `follow_up` responses; RpcClient.prompt() also accepts `streamingBehavior`.

— Named RPC methods and field, no example v0.99.0

13 **Theme color formats and appearance field** NEW 60

Adds `#rgb`, oklch(), and okhsl() color formats and an optional `appearance` field to theme files, plus theme.style(), `theme.colors`, and `theme.appearance` for extensions.

– Every format and API surface named, no usage shown v0.99.0

THINNER COVERAGE BELOW

14 **Claude Sonnet 5.5 support** NEW 50

Adds inherited Claude Sonnet 5.5 support for Anthropic with adaptive thinking and a 1M context window.

– Model, context size named, no config path given product docs

15 **Default theme changed to system** BREAKING 50

The default theme is changed to `system`; sessions that relied on the previous default dark/light appearance will now inherit the terminal palette.

– Clear before/after, no migration steps given product docs

16 **tool_search built-in extension** NEW 45

Adds a `tool_search` built-in extension that finds undeclared tools and declares them to the model at runtime.

– Thin one-line description with no config shown v0.99.0

17 **HTML export show/hide toggle** NEW 45

Adds a show/hide toggle (`H`) in HTML exports for custom messages marked `display: false`.

– Named key and field, minimal mechanism given v0.99.0

18 **provider_stream_event extension event** NEW 40

Adds the `provider_stream_event` extension event for observing parsed provider events before normalization.

– Named event, no usage detail given v0.99.0

19 **Model catalog type overlay** IMPROVED 40

Adds `types=chat,image,classifier` to pi.dev model catalog requests so remote refreshes overlay every supported model type.

– Named parameter, thin further detail v0.99.0

└──▷ BREAKING ON UPGRADE

! The inherited OpenAI Codex provider is renamed to OpenAI Codex (legacy); any configuration referencing it by name may need updating.

! The default OpenAI Codex model is changed to `gpt-6.1-sol`; sessions relying on the previous default model will now use GPT-6.1 Sol.

! The default theme is changed to `system`; sessions that relied on the previous default dark/light appearance will now inherit the terminal palette.

! `--no-extensions` now also disables built-in extensions (including the llama.cpp provider); previously these remained active.

! Built-in extensions and tools are now named `builtin:<name>` (e.g. `builtin:mcp`, `builtin:read`) in errors, diagnostics, RPC source info, and bug reports, replacing the previous `<inline:name>` / `<builtin:name>` naming.

! The default OpenAI Codex model is now `gpt-6.1-sol`; existing workflows that relied on the previous default Codex model will now use GPT-6.1 Sol unless a model is explicitly specified.

! `--no-extensions` now also disables the built-in extensions including the llama.cpp provider; previously built-ins were unaffected. Restore a specific built-in with `-e builtin:<name>`.

! Built-in extensions and tools are now named `builtin:<name>` (e.g. `builtin:mcp`, `builtin:read`) in errors, diagnostics, RPC source info, and bug reports — previously they appeared as `<inline:name>` and `<builtin:name>`.

v0.99.1 NOTES ↗ LATEST STABLE

Pi v0.99.1 adds GPT-6.1 Sol (`gpt-6.1-sol`) to OpenAI, Azure OpenAI, and OpenAI Codex providers and makes it the new default Codex model.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.99.1 https://github.com/earendil-
works/pi.git
# already have the repo? check out this version:
$ git checkout v0.99.1
```

└──▷ BREAKING ON UPGRADE

! The default OpenAI Codex model is now `gpt-6.1-sol`; existing workflows that relied on the previous default Codex model will now use GPT-6.1 Sol unless a model is explicitly specified.

v0.99.0 NOTES ↗ STABLE

Pi v0.99.0 adds MCP server support, codemode JavaScript sandbox, virtual/classifier models, and ChatGPT sign-in.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.99.0 https://github.com/earendil-
works/pi.git
# already have the repo? check out this version:
$ git checkout v0.99.0
```

└──▷ **USE IT**

Persist codemode as a default tool and cap its inline budget in your project config.

mcp.json

```
{
  "defaultTools": ["+codemode"],
  "codemode.mode": "auto",
  "codemode.inlineBudget": 4096
}
```

Adds `codemode` built-in extension that runs model-written JavaScript in a QuickJS sandbox; enable with `--tools` or `defaultTools` (e.g. `"+codemode"`) and configure with `codemode.mode` and `codemode.inlineBudget`.

Adds MCP server support over stdio or streamable HTTP with OAuth; configure globally or per project via `mcp.json`, register programmatically with pi.registerMcpServer(), and manage with `pi mcp add|remove|list|login|logout` or the `/mcp` slash command.

Adds `tool_search` built-in extension that finds undeclared tools and declares them to the model at runtime.

Adds extension tool API fields `exposure` (`direct`, `model-only`, `codemode`, `deferred`, or `hidden`), `namespace`, `annotations`, `outputSchema` with `structuredContent`, `isError` results, prepareLoadout(), and ctx.executeTool() for nested tool calls with `parentToolCallId` and `nestedCalls`.

Adds experimental virtual models: extensions register them with pi.registerVirtualModel() to route each request to a different physical model and thinking level; `/session` lists cost per physical model.

+13 more

└──▷ **BREAKING ON UPGRADE**

! `--no-extensions` now also disables the built-in extensions including the llama.cpp provider; previously built-ins were unaffected. Restore a specific built-in with `-e builtin:<name>`.

! Built-in extensions and tools are now named `builtin:<name>` (e.g. `builtin:mcp`, `builtin:read`) in errors, diagnostics, RPC source info, and bug

reports — previously they appeared as `<inline:name>` and `<builtin:name>`.

WAS THIS USEFUL? 👍 👎

Superset

Feedback ⓘ desktop-v1.32.0 NOTES ↗ RANK

Superset is an agentic IDE to orchestrate 100+ coding agents in parallel. Run any agent with your own subscription.

> Superset desktop v1.32.0 adds CLI-driven cloud and GitHub App setup, recoverable project deletion, cloud workspace automations, and new page capabilities including reporting and agent-generated pages.

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

01 **`superset stop --terminals` CLI subcommand** NEW 75

A new `superset stop --terminals` CLI subcommand ends the terminal daemon, letting users cleanly stop the background terminal daemon when shutting down a development session.

Stop the background terminal daemon cleanly when shutting down a development session.

```
$ superset stop --terminals
```

— Exact runnable CLI command given desktop-v1.32.0

THINNER COVERAGE BELOW

02 **Recoverable project deletion** IMPROVED 45

Project deletion is now recoverable — deleted projects can be restored within 30 days.

— Concrete 30-day window but no restore mechanism named desktop-v1.32.0

03 **CLI setup for cloud environments and GitHub App** NEW 42

The CLI now enables coworkers to set up their own cloud environment, and enables installing the organization's GitHub App, both from the CLI.

— Named as CLI-driven but no exact commands given desktop-v1.32.0

04 **Public page reporting and agent-generated pages** NEW 35

Users can now report a public page, and staff can take a reported page down; separately, agents now offer a page when their output has a reader.

— Names two page behaviors but no UI path or mechanism desktop-v1.32.0

05 **Reorganized desktop settings UI** IMPROVED 23

Browser and file editor settings are now organized in the desktop UI.

— Vague UI reorganization with no navigation detail desktop-v1.32.0

06 **Automations in cloud workspaces** NEW 20

Automations can now be run in cloud workspaces.

— Bare capability statement with no mechanism desktop-v1.32.0

desktop-v1.32.0 NOTES ↗ LATEST STABLE

Superset desktop v1.32.0 adds CLI-driven cloud setup, GitHub App install, recoverable project deletion, and new automation and pages capabilities.

└──▷ **GET THIS VERSION**

```
$ git clone --branch desktop-v1.32.0 https://github.com/superset-sh/superset.git
# already have the repo? check out this version:
$ git checkout desktop-v1.32.0
```

└──▷ **TRY IT**

Stop the background terminal daemon cleanly when shutting down a development session.

```
$ superset stop --terminals
```

Adds `superset stop --terminals` CLI subcommand to end the terminal daemon.

Enables coworkers to set up their own cloud environment from the CLI.

Enables installing the organization's GitHub App from the CLI.

Adds recoverable project deletion — deleted projects can be restored within 30 days.

Enables running automations in cloud workspaces.

+3 more

WAS THIS USEFUL? 👍 👎

Alibaba Qwen Code

Feedback ⓘ sdk-typescript-v0.1.17

NOTES ↗ RANK

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

> Qwen Code's SDK v0.1.17 adds multi-modal input, session lifecycle management (resume, continue, export), an experimental LSP, a Zed extension, and a concurrent batch runner, alongside several breaking changes to settings naming and legacy tools.

└──▷ **WHAT SHIPPED · 17 FEATURES** most completely described first ? what's the number?

01 **Session resume, continue and export in SDK** NEW 80

Adds `resume` and `continue` options plus extended `authType` support to the TypeScript SDK, and supports resuming sessions by session ID via the `query` interface. A new `export` command exports session history to markdown and HTML output formats.

— Names flags and command but no exact syntax shown sdk-typescript-v0.1.17

02 **New config keys: settings.env and contextWindowSize** NEW 80

Adds a `settings.env` field for configuring environment variables directly in settings — useful for injecting API keys or proxy settings without exporting them to the shell — and a `contextWindowSize` configuration key.

Set environment variables for a session without exporting them to the shell — useful for injecting API keys or proxy settings via config.

json

```
{
  "env": {
    "HTTP_PROXY": "http://proxy.corp.example.com:8080",
    "MY_API_KEY": "sk-..."
  }
}
```

— Named config keys with a runnable example config sdk-typescript-v0.1.17

03 **Prompt-cache reuse for Anthropic content generator** IMPROVED 65

Adds `cache_control` for system and last user text message in the Anthropic content generator to improve prompt-cache reuse.

– Names field and gives purpose, no config path sdk-typescript-v0.1.17

04 **Managed memory gated by enableManagedAutoMemory** IMPROVED

60

Managed memory availability now respects the `memory.enableManagedAutoMemory` setting.

– Exact config key given, minimal behavioural detail sdk-typescript-v0.1.17

05 **Settings keys renamed from disable* to enable*** BREAKING

60

Negative settings keys are renamed to positive naming — `disable*` becomes `enable*` — so any config referencing old `disable*` keys will stop working.

– Explains impact but no full list of renamed keys sdk-typescript-v0.1.17

06 **read_many_files tool removed** BREAKING

60

The `read_many_files` tool is removed; `readManyFiles` is now a utility for user @`-commands` only.

– Names old and new surfaces and their scope sdk-typescript-v0.1.17

THINNER COVERAGE BELOW

07 **FORK_MODE for ProcessTransport Electron IPC** NEW

55

Adds `FORK_MODE` support to `ProcessTransport` for Electron IPC integration.

– Names exact flag and transport class sdk-typescript-v0.1.17

08 **Multi-modal input across content generators** NEW

50

All content generators now accept multi-modal input — image, PDF, and audio — rather than text only.

– States scope but no mechanism or usage example sdk-typescript-v0.1.17

09 **Logfile-first debug logger** IMPROVED

50

Debug mode output is refactored to route console calls to a logfile-first `debugLogger`.

– Names the mechanism but no usage path sdk-typescript-v0.1.17

10 **extra_body support for OpenAI-compatible providers** NEW

40

Adds `extra_body` support for OpenAI-compatible providers.

– Bare name, no mechanism or usage detail sdk-typescript-v0.1.17

11 **Agent Skills stable with symlink support** IMPROVED 40

Promotes Agent Skills from experimental to stable, and adds symlink support for the skill manager.

— Notes status change and one added capability, no detail sdk-typescript-v0.1.17

12 **Concurrent batch CLI runner** NEW 35

Adds a concurrent runner for batch CLI execution.

— One-line description with no configuration detail sdk-typescript-v0.1.17

13 **Experimental LSP support** NEW 35

Adds experimental LSP support for code intelligence.

— Names the surface but not how to enable it sdk-typescript-v0.1.17

14 **MCP tool progress updates in TUI and SDK** IMPROVED 35

Adds MCP tool progress update support in TUI and SDK mode.

— Thin description, no mechanism given sdk-typescript-v0.1.17

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

Adds a Zed extension for the Qwen Code agent server.

— Integration named but no setup instructions sdk-typescript-v0.1.17

16 **Smart Edit tool and ClearcutLogger removed** BREAKING 30

The Smart Edit tool and ClearcutLogger are removed.

— Names removed components, no migration guidance sdk-typescript-v0.1.17

17 **Large paste placeholder and macOS enter-submit fix** IMPROVED 25

Adds a large paste placeholder and improves enter-submit behaviour on macOS.

— Minor UI tweak with minimal detail sdk-typescript-v0.1.17

└──▷ **BREAKING ON UPGRADE**

! Negative settings keys renamed to positive naming (`disable*` becomes `enable*`) — any config referencing old `disable*` keys will stop working.

! The `read_many_files` tool is removed; `readManyFiles` is now a utility for user @`-commands` only.

! The Smart Edit tool and ClearcutLogger are removed.

sdk-typescript-v0.1.17 NOTES ↗ LATEST STABLE

Qwen Code SDK v0.1.17 adds multi-modal input, session resume, `settings.env`, export command, LSP, Zed extension, and concurrent batch runner.

└──▷ **GET THIS VERSION**

```
$ git clone --branch sdk-typescript-v0.1.17
https://github.com/QwenLM/qwen-code.git
# already have the repo? check out this version:
$ git checkout sdk-typescript-v0.1.17
```

└──▷ **USE IT**

Set environment variables for a session without exporting them to the shell — useful for injecting API keys or proxy settings via config.

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

Adds `settings.env` field for configuring environment variables directly in settings.

Adds `resume` and `continue` options plus extended `authType` support to the TypeScript SDK.

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

Adds support for resuming sessions by session ID via the `query` interface.

Adds `extra_body` support for OpenAI-compatible providers.

+13 more

└──▷ **BREAKING ON UPGRADE**

- ! Negative settings keys renamed to positive naming (`disable*` becomes `enable*`) — any config referencing old `disable*` keys will stop working.
- ! The `read_many_files` tool is removed; `readManyFiles` is now a utility for user @ `-commands` only.
- ! The Smart Edit tool and ClearcutLogger are removed.

WAS THIS USEFUL?  

Sourcegraph Amp

Feedback ⓘ Plaid Speed (2026-09-29) BLOG ↗

RANK

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

> Amp expanded plugin confirmation dialogs with new interactive field types and introduced a Plaid speed tier for GPT-6 Astra modes offering up to 6× faster inference at a premium cost.

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

01 Editable, image, and link fields in plugin confirmation dialogs

NEW 80

Plugin confirmation dialogs gain editable text fields (defaulting to a supplied value and requiring non-empty input to confirm), a `type: 'image'` field option showing a read-only HTTP/HTTPS image preview, and a `type: 'link'` field option rendering a labelled button that opens an HTTP/HTTPS URL in the browser while keeping the dialog open.

Add an approval link to a confirmation dialog so reviewers can open an external approval page before confirming or cancelling.

```json
{
  type: 'link',
  label: 'Review in Approvals',
  value: 'https://approvals.example.com/requests/42'
}
```

— Names exact field types and behaviour with a config example product docs

02 Plaid speed tier for GPT-6 Astra modes NEW 70

Amp adds a Plaid speed tier for modes using GPT-6 Astra (e.g., High mode), running up to 6× faster than standard at 6× cost per token. It's toggled via the speed switch in the thread dialog, past Fast, and can be turned on or off for an existing thread, mirroring the existing Fast mode toggle behaviour.

.

— Clear mechanism and numbers, but only a UI toggle path Plaid Speed (2026-09-29)

Plaid Speed (2026-09-29) BLOG ↗ LATEST STABLE

> Amp adds Plaid speed tier for GPT-6 Astra modes, delivering up to 6× faster inference via OpenAI's ultrafast tier.
>
> Adds Plaid speed tier for modes using GPT-6 Astra (e.g., High mode), running up to 6× faster than standard at 6× cost per token, toggled via the speed switch in the thread dialog past Fast to Plaid.
>
> Plaid speed can be toggled on or off for an existing thread, mirroring the existing Fast mode toggle behaviour.

WAS THIS USEFUL?  

Warp

Feedback ⓘ announced · Sep 29, 2026 BLOG ↗ RANK

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

> Warp added Sign in with ChatGPT authentication, letting users apply their OpenAI subscription quota to AI requests in the terminal and unlocking new GPT-6 models.

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

01 **Sign in with ChatGPT and GPT-6 model access** NEW 65

Warp Terminal and Warp Agent CLI now support Sign in with ChatGPT, letting users authenticate with their OpenAI account and apply their subscription's Work and Codex usage toward AI requests in Warp. Via this ChatGPT sign-in integration, Warp Agent also gains support for GPT-6 models Astra, Sol, Terra, and Luna.

Sign in to Warp with your ChatGPT account to use your existing OpenAI subscription quota for AI requests in the terminal.

> 📍 In Warp Terminal, open the sign-in screen and select Sign in with ChatGPT, then authenticate with your OpenAI account credentials.

— Names auth flow and four model names but no config detail launch-20260929-07e5a0d4

> announced · Sep 29, 2026 BLOG ↗
>
> Warp Terminal and Warp Agent CLI now support Sign in with ChatGPT, letting users authenticate via their OpenAI account.
>
> └──▷ **HOW TO FIND IT**
>
> Sign in to Warp with your ChatGPT account to use your existing OpenAI subscription quota for AI requests in the terminal.
>
> > 📍 In Warp Terminal, open the sign-in screen and select Sign in with ChatGPT, then authenticate with your OpenAI account credentials.
>
> Adds Sign in with ChatGPT authentication to Warp Terminal and Warp Agent CLI, allowing users to authenticate with their ChatGPT account and apply their subscription's Work and Codex usage toward AI requests in Warp.
>
> Warp Agent now supports GPT-6 models Astra, Sol, Terra, and Luna via the ChatGPT sign-in integration.

WAS THIS USEFUL?  

Amazon Kiro

Feedback ⓘ

CLI: Power Management, V3 Output Styles, and Fullscreen Scroll Speed: Install and uninst

BLOG ↗ RANK

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

> Kiro added Claude Opus 5.5 with a 1M context window to Kiro Web, while CLI 2.25 shipped Power management from chat and terminal, a SessionEnd teardown Hook, V3 output styles, and tunable fullscreen scrolling.

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

01 Power management from chat and CLI NEW

90

Kiro CLI 2.25 adds `/powers install <name|path>` and `/powers uninstall <name>` slash commands to install or remove Powers within a running V3 chat session, with an immediate refresh so the session can use the updated set. It also adds `kiro-cli powers install <name|path>` and `kiro-cli powers uninstall <name>` subcommands to manage Powers without starting a chat session, useful for CI or setup scripts.

Install a Power by name from the catalog without launching a chat session, useful in CI or setup scripts.

```
$ kiro-cli powers install <name|path>
```

Install a Power mid-session from the local catalog or a local directory path so it is available immediately in the current V3 chat.

📍 `/powers` install <name|path>

— Names all four runnable commands and their behavior

CLI: Power Management, V3 Output Styles…

02 Fullscreen display settings and scroll speed IMPROVED

75

Kiro CLI adds Full Screen display settings under `/settings display`, with a Scroll speed option accepting 1, 2, or 3 rows per mouse-wheel event (default 2), consolidated alongside the existing Start fullscreen setting in the same screen.

Tune fullscreen scroll speed to 1 row per wheel event for precise transcript navigation in a V3 session.

📍 Open `/settings display`, select `Full Screen`, and set `Scroll speed` to 1.

— Exact settings path and concrete scroll-speed values given

CLI: Power Management, V3 Output Styles…

03 **Claude Opus 5.5 in Kiro Web** NEW 60

Kiro Web now supports switching to Claude Opus 5.5, giving large agentic coding sessions access to a 1M context window. Users refresh their browser, open the model selector, and choose Claude Opus 5.5.

Switch to Claude Opus 5.5 in the Kiro Web interface to take advantage of the 1M context window for large agentic coding sessions.

📍 In Kiro Web, refresh your browser, then open the model selector and choose Claude Opus 5.5.

— Names the model and context size but no other mechanism launch-20260928-0111ecbc

THINNER COVERAGE BELOW

04 **SessionEnd Hook for teardown automation** NEW 55

A new `SessionEnd` Hook trigger runs automation when a V3 session shuts down, enabling reliable teardown and cleanup tasks.

— Names the trigger but gives no setup command CLI: Power Management, V3 Output Styles…

05 **V3 output styles via /settings** NEW 50

Kiro CLI adds `/settings` in a V3 session to choose V3 output styles, including styles supplied by upgraded agents so each agent can offer formats suited to its work.

— Names the command but not the style options themselves

CLI: Power Management, V3 Output Styles…

CLI: Power Management, V3 Output Styles, and Fullscreen Scroll Speed: Install and unin

BLOG ↗ LATEST STABLE

Kiro CLI 2.25 adds in-chat and CLI Power management, V3 output styles, fullscreen scroll tuning, and a SessionEnd Hook.

└──▷ **TRY IT**

Install a Power by name from the catalog without launching a chat session, useful in CI or setup scripts.

```
$ kiro-cli powers install <name|path>
```

Install a Power mid-session from the local catalog or a local directory path so it is available immediately in the current V3 chat.

📍 `/powers` install <name|path>

Tune fullscreen scroll speed to 1 row per wheel event for precise transcript navigation in a V3 session.

📍 Open `/settings display`, select `Full Screen`, and set `Scroll speed` to 1.

Adds `/powers install <name|path>` and `/powers uninstall <name>` slash commands to install or remove Powers from within a running V3 chat session, with immediate refresh so the current session can use the updated set.

Adds `kiro-cli powers install <name|path>` and `kiro-cli powers uninstall <name>` subcommands to manage Powers without starting a chat session.

Adds a `SessionEnd` Hook trigger to run automation when a V3 session shuts down, enabling reliable teardown and cleanup tasks.

Adds `/settings` in a V3 session to choose V3 output styles, including styles supplied by upgraded agents so each agent can offer formats suited to its work.

Adds `Full Screen` display settings under `/settings display`, with a `Scroll speed` option (1, 2, or 3 rows per mouse-wheel event; default 2) and the existing `Start fullscreen` setting consolidated in the same screen.

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

AutoGPT

Feedback ⓘ autogpt-platform-beta-v0.8.2 NOTES ↗ RANK

AutoGPT automates multi-step workflows with continuous AI agents that users can build and deploy.

v0.8.2 introduces AutoPilot action-gating modes (Ask First, Auto, Unsupervised), prompt-injection holds on outside reads, MCP server trust/effect-map evaluation, scoped approval rules, a credential swap proxy for E2B sandboxes, AnySearch blocks, and async Expert hiring.

└──▷ GET THIS VERSION

```
$ git clone --branch autogpt-platform-beta-v0.8.2
https://github.com/Significant-Gravitas/AutoGPT.git
# already have the repo? check out this version:
$ git checkout autogpt-platform-beta-v0.8.2
```

Adds AutoPilot action-gating modes — Ask First, Auto, and Unsupervised — controlling whether the agent pauses for human approval before acting.

Adds approval rules scoped to one chat, one Expert, or the whole team, with `Approve from now on` and `Let Otto judge from now on` options on bare-tool cards.

Adds a credential swap proxy that all E2E sandbox (E2B) boxes egress through, centralising outbound credential handling.

Adds `AnySearch` search, parallel-search, and extract blocks for web-search workflows.

AutoPilot pauses workflows before irreversible actions and continues processing other queued cards while a held call awaits review.

+14 more

NEW INTEGRATION PostHog MCP integration

NEW INTEGRATION OpenSEO MCP integration

NEW SUPPORTED PLATFORM xAI Grok 4.7 LLM

NEW SUPPORTED PLATFORM GLM-5.3 LLM

NEW SUPPORTED PLATFORM Claude Opus 4.8 LLM

NEW SUPPORTED PLATFORM Xiaomi MiMo-V2.6-Pro and MiMo-V2.6-Flash LLMs

NEW SUPPORTED PLATFORM Claude Sonnet 5.5 LLM (new standard-tier default)

UPDATED SUPPORTED PLATFORM Fast/standard default LLM promoted to gpt-6-luna

└──▷ **BREAKING ON UPGRADE**

! Marketplace listings now require a short CTA-style `sub_heading` field; existing listings are backfilled by a migration script but any tooling that creates listings programmatically must supply this field.

WAS THIS USEFUL?  

OpenClaw

Feedback ⓘ v2026.9.7 NOTES ↗ CODE ↗ RANK

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

> OpenClaw v2026.9.7 adds a new OpenAI Agents API plugin, Sign in with ChatGPT, worktree sessions, a `decision_evaluate` tool, an opt-in Gemini Interactions backend and Telnyx as a model provider, alongside a broad set of operator controls, session and privacy features, Apple/voice/chat-media improvements, and performance speedups.

└──▷ **WHAT SHIPPED · 26 FEATURES** most completely described first ? what's the number?

01 Crabbox run with protected model credential NEW 80

Adds `openclaw crabbox run --model` subcommand to run applications inside a Crabbox lease with a protected model credential that never leaves its original host.

Run a sandboxed application inside a Crabbox lease with a model credential that stays on the originating host — useful for isolating sensitive workloads from the client environment.

```
$ openclaw crabbox run --model <model>
```

— Runnable command with concrete security mechanism described. v2026.9.7

02 Configured-model default for new sessions NEW 75

Adds the `gateway.controlUi.newSessionModelDefaults: "configured"` config key to start fresh sessions from the agent's configured model, runtime, and thinking defaults instead of last-used choices; the default remains `"last-used"`.

Pin fresh sessions to the agent's configured model instead of inheriting whatever the last user picked — useful in regulated environments where model drift is a compliance risk.

```yaml
gateway:
  controlUi:
    newSessionModelDefaults: "configured"
```

— Names exact config key, default value, and behaviour change. v2026.9.7

03 OpenAI Agents API plugin NEW 75

Adds a new OpenAI Agents API plugin supporting streamed replies, steering, live web search, OpenClaw tools, attachments, hosted files, preserved tool history, persona/workspace context, self-hosted skill discovery, and accurate token usage.

— Rich capability list but no install command or config key. v2026.9.7

04 **Inline chat approval workflow** NEW 75

Adds operator ability to hand the agent an API key, config change, or skill edit directly in chat, with `/approve <id>` completing OpenClaw change approvals inline with Allow once and Deny buttons.

— Names exact command and UI controls for the workflow. v2026.9.7

05 **Unmentioned follow-ups in bot threads** NEW 70

Adds `requireMentionInBotThreads: false` config key to accept unmentioned follow-ups in bot-created threads across Discord, Slack, Telegram, Matrix, Mattermost, Microsoft Teams, iMessage, Feishu, and other supported channels.

Allow users in Slack or Discord threads to reply without @`-mentioning` the bot, so follow-up questions in bot-created threads flow naturally without forcing a mention every time.

yaml

```yaml
requireMentionInBotThreads: false
```

— Exact config key and full channel list with runnable example. v2026.9.7

06 **decision_evaluate tool for structured decisions** NEW 65

Adds the `decision_evaluate` tool for Boolean, Choice, and Score questions against an agent's configured decision model, with provider capabilities listed in `models.list`.

— Named tool and related surface, but no usage example. v2026.9.7

07 **Portal tool for scoped previews** NEW 65

Adds the restricted `portal` tool to let collaborators open session-scoped interactive previews from dedicated worker attachments without global host-port access.

— Named tool with clear scope and access-control mechanism. v2026.9.7

08 **Incognito session privacy controls** NEW 65

Adds Incognito session controls: sessions now expire after 24 hours, agent turns are kept out of task records and logs, and terminals stop on session reset.

— Concrete behaviour and 24-hour limit, no toggle command. v2026.9.7

09 **Ultrafast mode opt-in for Codex plugin** NEW 60

Adds `plugins.entries.codex.config.appServer.enableUltrafast` config opt-in for Ultrafast mode, active only when the selected model advertises support for it.

— Names config key but no example or command shown. v2026.9.7

10 **Worktree sessions for parallel chats** NEW 60

Adds worktree sessions — start a new chat in an isolated managed worktree from web, iOS, or Android so parallel sessions on one repository no longer collide.

— Explains behaviour and platforms, no exact command. v2026.9.7

11 **Gemini Interactions API backend** NEW 60

Adds an opt-in `google-interactions` API backend for Gemini Interactions text, image, reasoning, and tool-call workflows.

— Named backend and workflow types, no config example. v2026.9.7

12 **Role-based model policy limits** NEW 60

Adds role model limits: administrators can assign named operator roles a model policy enforced by native agents, Visitor Access, model pickers, and plugin completions.

— Names enforcement surfaces but no config key shown. v2026.9.7

13 **Native Apple chat enhancements** IMPROVED 60

Extends native Apple (macOS and iOS) chat with message timestamps, model identity per reply, agent identities, file and audio attachments, and on macOS a command palette plus web sidebar roster defaults and view options.

— Lists several concrete additions but no exact steps. v2026.9.7

14 **Anthropic Claude Opus 5.5 support** IMPROVED 60

Improves Anthropic support: bare `opus` and new Anthropic API, CLI, and media setup now select Claude Opus 5.5, and thinking-display preferences reach Anthropic transports.

— Names model version and setup surfaces, no exact command. v2026.9.7

THINNER COVERAGE BELOW

15 **Desktop audio streaming** NEW 55

Adds desktop audio streaming from the docked desktop panel or standalone desktop view, with isolated per-application audio on managed Linux desktops.

— Names UI surfaces and isolation mechanism, no config path. v2026.9.7

16 **Unified voice agent context** IMPROVED 55

Extends voice (Talk, Discord, FaceTime, and Twilio) so all calls share one agent context, enabling voice answers about other sessions to delegate to the agent rather than deny.

— Describes before/after behaviour but no configuration detail. v2026.9.7

17 **Chat media viewing improvements** NEW 50

Adds inline PDF preview, navigation between expanded videos within a turn, and copy-as-Markdown for your own messages.

— Names three UI additions, no exact navigation path. v2026.9.7

18 **Duplicate-safe Google Meet participation** NEW 45

Adds experimental duplicate-safe Google Meet participation with retained caption context and speaker provenance.

— Describes behaviour but marked experimental, no setup path. v2026.9.7

19 **Activity Sessions Today pulse** NEW 45

Adds an Activity Sessions tab with a Today pulse of hourly activity and counts that follow current filters.

— Names UI tab, gives a clear starting point. v2026.9.7

20 **Plugin browsing and discovery upgrades** IMPROVED 45

Improves plugin browsing with declared capabilities, setup guides, tool-input detail previews, curated categories, and a new Computer use shelf.

— Lists UI additions without exact navigation or config. v2026.9.7

21 **Model pricing and catalog updates** IMPROVED 45

Improves model pricing: each route is now priced by whoever bills it, and a smaller v2 hosted catalog with ordered model recommendations is downloaded.

— Describes change but no config key or numbers. v2026.9.7

22 **CLI and session performance improvements** IMPROVED 45

Speeds up `openclaw --help` and skips resolving channel options at startup for commands that do not need them; also speeds up session lists, chat history pages, archived session search, and transcript appends.

— Names affected commands and views, no measured speedup. v2026.9.7

23 **Sign in with ChatGPT authentication** NEW 40

Adds Sign in with ChatGPT (Beta) as an OpenAI authentication choice alongside Codex login and API keys.

— Names the option but no setup steps given. v2026.9.7

24 **Telnyx model provider** NEW 40

Adds Telnyx as an official external model provider installable through normal onboarding.

— Names provider but no onboarding detail given. v2026.9.7

25 **Subagent delegation guidance** IMPROVED 40

Adds `sessions_spawn` subagent guidance clarifying when research is worth delegating and how to follow up on a stalled child by checking execution and delivery status.

— Named tool guidance but no command or example. v2026.9.7

26 **Disable client file and image uploads** NEW 30

Adds operator control to disable client file and image uploads to the Gateway.

— Thin description with no named toggle or config key. v2026.9.7

v2026.9.7 NOTES ↗ CODE ↗ LATEST STABLE

OpenClaw v2026.9.7 adds OpenAI Agents API plugin, Sign in with ChatGPT, worktree sessions, decision_evaluate tool, Gemini Interactions backend, and Telnyx provider.

└──▷ GET THIS VERSION

```
$ git clone --branch v2026.9.7
https://github.com/openclaw/openclaw.git
# already have the repo? check out this version:
$ git checkout v2026.9.7
```

└──▷ USE IT

Pin fresh sessions to the agent's configured model instead of inheriting whatever the last user picked — useful in regulated environments where model drift is a compliance risk.

```yaml
gateway:
  controlUi:
    newSessionModelDefaults: "configured"
```

Allow users in Slack or Discord threads to reply without @`-mentioning` the bot, so follow-up questions in bot-created threads flow naturally without forcing a mention every time.

```yaml
requireMentionInBotThreads: false
```

Run a sandboxed application inside a Crabbox lease with a model credential that stays on the originating host — useful for isolating sensitive workloads from the client environment.

```
$ openclaw crabbox run --model <model>
```

Adds `gateway.controlUi.newSessionModelDefaults: "configured"` config key to start fresh sessions from the agent's configured model, runtime, and thinking defaults instead of last-used choices (default remains `"last-used"`).

Adds `requireMentionInBotThreads: false` config key to accept unmentioned follow-ups in bot-created threads across Discord, Slack, Telegram, Matrix, Mattermost, Microsoft Teams, iMessage, Feishu, and other supported channels.

Adds `plugins.entries.codex.config.appServer.enableUltrafast` config opt-in for Ultrafast mode, active only when the selected model advertises it.

Adds `openclaw crabbox run --model` subcommand to run applications inside a Crabbox lease with a protected model credential that never leaves its original host.

Adds the `decision_evaluate` tool for Boolean, Choice, and Score questions against an agent's configured decision model, with provider capabilities listed in `models.list`.

+22 more

WAS THIS USEFUL? 👍 👎

PydanticAI

Feedback ⓘ v2.52.0 NOTES ↗ RANK

PydanticAI is a framework that builds type-safe Python agents with dependency injection, model integrations, tools, and structured outputs.

> PydanticAI v2.52.0 introduces a unified `ctx.workspace` API for harness tools alongside new sandbox backends (SpritesSandbox, E2BSandbox, SSHWorkspace, BubblewrapSandbox), a wave of built-in clai2 CLI plugins, and realtime model updates spanning Azure Voice Live, OpenAI GPT-Live, and Gemini 3.x Live, plus new Claude Sonnet 5.5 and gpt-6.1-sol model support.

└──▷ **WHAT SHIPPED · 17 FEATURES** most completely described first ? what's the number?

01 Async tool call and prompt cache profile settings NEW

88

Adds `async_tool_call_mode` profile setting and a shared `async_tool_calls` setting to control model realtime async tool calls, plus a `default_cache_retention` profile key and a prompt_cache_outlook() cold-window helper for prompt caching.

Use the new async tool call profile settings to control how realtime sessions handle parallel tool calls.

python

```python
from pydantic_ai import Agent
from pydantic_ai.profiles import ModelProfile

profile = ModelProfile(
    async_tool_call_mode='sequential',
    async_tool_calls=True,
)
agent = Agent('openai:gpt-6.1-sol', model_profile=profile)
```

— Named settings with a runnable code example v2.52.0

02 clai2 --agent flag for plugin-free chat NEW

75

Adds `clai2 --agent MODULE:ATTR` flag to start a chat session with an existing Agent with plugins disabled.

Chat with an existing Agent from the terminal via clai2, bypassing all plugins, for a clean one-off session.

```
$ uvx pydantic-clai2 --agent my_package.agents:research_agent
```

– Named flag with a runnable command example v2.52.0

03 **Built-in clai2 plugins for third-party services** NEW

70

clai2 ships built-in plugins for `github` (settings menu, token in `/keys`), `slack` (settings menu, `/keys` user token, or browser sign-in), `google_workspace` (backed by harness `GoogleWorkspace`), `notion` (settings menu, key in `/keys`), `day_ai` (settings menu, `/keys` or browser sign-in), `ordinal` (token or keyring-backed browser sign-in), `pylon` (settings menu, named `/keys` entry), and `logfire_mcp` (`/keys`, `LOGFIRE_API_KEY`, keyring-backed OAuth).

– Every plugin and its auth surface named, no command v2.52.0

04 **OpenAI GPT-Live and Responses realtime updates** NEW

68

Adds a `_process_provider_details` hook to `OpenAIResponsesModel`, supports browser WebRTC on OpenAI GPT-Live via `answer_webrtc_offer` and a sideband session, emits identified response, turn, and input lifecycle events from the OpenAI-protocol realtime connection, and maps the shared `thinking` and `parallel_tool_calls` settings to the GPT-Live backend model.

– Several named hooks and events, no example v2.52.0

05 **AnthropicModel max_tokens defaults raised** IMPROVED

65

Raises `AnthropicModel`'s default `max_tokens` from 4096 to 16384 on Claude Sonnet 4.5 and later, and now defaults `max_tokens` to the model's maximum output, streaming such requests behind the scenes.

– Clear before/after numbers, no code sample v2.52.0

06 **Azure AI Voice Live realtime settings** NEW

63

Adds `azure_voice_live_voice` to choose an Azure voice for Azure AI Voice Live sessions, and applies `thinking`, `openai_turn_detection`, `openai_input_noise_reduction`, and a new `azure_voice_live_temperature` setting on Azure AI Voice Live.

– Multiple named settings, no usage example v2.52.0

07 **Unified ctx.workspace API for harness tools** NEW

62

Adds a `ctx.workspace` API so harness tools (`Coder`, `FileSystem`, `Shell`) share one unified interface for files and commands, whether running locally, in a sandbox, or with durable execution support.

– Names the API and tools but no usage example v2.52.0

THINNER COVERAGE BELOW

08 **New sandbox and workspace backends for harness** NEW 58

Adds `SpritesSandbox`, a workspace that runs `Coder`, `Shell`, `FileSystem`, and every other harness capability inside a Fly.io Sprite; `E2BSandbox`, which runs the same harness capabilities inside an E2B sandbox; and `SSHWorkspace` and `BubblewrapSandbox` as additional new harness workspace capabilities.

— Names four new backends but no setup instructions v2.52.0

09 **Gemini 3.x Live realtime enhancements** NEW 58

Seeds retained user audio on Gemini `gemini-3.1-flash-live-preview` and `gemini-3.8-live`, seeds tool calls and results as native function parts on Gemini `gemini-3.8-live`, and delivers images and text files returned from tools on Gemini 3.x Live.

— Named models and behaviors, no example v2.52.0

10 **New model support: Claude Sonnet 5.5, gpt-6.1-sol** NEW 58

Adds Claude Sonnet 5.5 (`claude-sonnet-5-5`) model support, OpenAI `gpt-6.1-sol` model support, and allows image output on `gpt-6-astra`.

— Model identifiers named, no usage detail v2.52.0

11 **Tool-call arguments shown in clai2 headers** IMPROVED 57

Shows tool-call arguments in `pydantic-clai2` tool headers, clipped by the `display.tool_arg_chars` setting.

— Names config key, no example of use v2.52.0

12 **ModalSandboxBackend replaces ModalSandboxSession** BREAKING 55

`ModalSandboxBackend` replaces `ModalSandboxSession`; existing code referencing `ModalSandboxSession` will break and must be updated.

— Clear migration target named, no code sample v2.52.0

13 **typesafe extra rejects unavailable providers** NEW 50

Adds a `typesafe` extra to `pydantic-clai2` that rejects models with a missing provider SDK at `/model` time.

— Names the extra and check point, no example v2.52.0

14 **OpenRouter quantization and sort options** NEW 50

Accepts `mxfp4`, `nvfp4`, and `mxfp8` OpenRouter `quantizations` values and the `exacto` `sort` value.

— Named values but no usage context v2.52.0

15 **pydantic-ai-harness and pydantic-clai2 versioned together**

IMPROVED 50

Ships `pydantic-ai-harness` and `pydantic-clai2` `0.52.0` together with every Pydantic AI release; `pydantic-clai2` is available via `uvx pydantic-clai2`.

— Names packages and install command, thin detail v2.52.0

16 **SubAgents no longer auto-loads agent files** BREAKING 45

`SubAgents` no longer loads agent files by default, and the `inherit_tools` parameter is now deprecated.

— Behavior change stated but no migration path given v2.52.0

17 **resolve_prompt_cache_retention renamed** BREAKING 45

resolve_prompt_cache_retention() is renamed to resolve_cache_retention(); callers must update their code.

— Exact rename named, no code migration shown v2.52.0

└──▷ **BREAKING ON UPGRADE**

! `ModalSandboxBackend` replaces `ModalSandboxSession`; code referencing `ModalSandboxSession` will break.

! `SubAgents` no longer loads agent files by default; `inherit_tools` is deprecated.

! resolve_prompt_cache_retention() is renamed to resolve_cache_retention(); callers must update.

v2.52.0 NOTES ↗ LATEST STABLE

PydanticAI v2.52.0 adds workspaces, SpritesSandbox, E2BSandbox, SSHWorkspace, BubblewrapSandbox, clai2 plugins, and new model/realtime APIs.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v2.52.0 https://github.com/pydantic/pydantic-ai.git
# already have the repo? check out this version:
$ git checkout v2.52.0
```

└──▷ **TRY IT**

Chat with an existing Agent from the terminal via clai2, bypassing all plugins, for a clean one-off session.

```
$ uvx pydantic-clai2 --agent my_package.agents:research_agent
```

Use the new async tool call profile settings to control how realtime sessions handle parallel tool calls.

```python
from pydantic_ai import Agent
from pydantic_ai.profiles import ModelProfile

profile = ModelProfile(
    async_tool_call_mode='sequential',
    async_tool_calls=True,
)
agent = Agent('openai:gpt-6.1-sol', model_profile=profile)
```

Adds `ctx.workspace` API so harness tools (`Coder`, `FileSystem`, `Shell`) share one unified interface for files and commands, whether running locally, in a sandbox, or with durable execution support.

Adds `SpritesSandbox` workspace that runs `Coder`, `Shell`, `FileSystem`, and every other harness capability inside a Fly.io Sprite.

Adds `E2BSandbox` workspace that runs `Coder`, `Shell`, `FileSystem`, and every other harness capability inside an E2B sandbox.

Adds `SSHWorkspace` and `BubblewrapSandbox` as new harness workspace capabilities.

Adds `clai2 --agent MODULE:ATTR` flag to start a chat session with an existing `Agent` with plugins disabled.

+27 more

└──▷ **BREAKING ON UPGRADE**

- ! `ModalSandboxBackend` replaces `ModalSandboxSession`; code referencing `ModalSandboxSession` will break.
- ! `SubAgents` no longer loads agent files by default; `inherit_tools` is deprecated.
- ! resolve_prompt_cache_retention() is renamed to resolve_cache_retention(); callers must update.

WAS THIS USEFUL? 👍 👎

Feynman

Feedback ⓘ v0.5.15 NOTES ↗ RANK

Feynman is a personal AI research assistant that searches papers and the web, builds literature reviews, and produces cited research reports.

> Feynman ships MCP server support via an upgraded, pinned Pi dependency, a fuller set of `feynman model` CLI commands for picking and diagnosing research models, and a fix for OpenAlex search failures caused by rate limiting.

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

01 **Model selection commands in the CLI** NEW 95

Feynman adds `feynman model` to list available models, `feynman model set [provider/model]` to set the default approved research model for all sessions (opening an interactive picker when run without an argument), and `feynman model help` to list all model subcommands (unknown subcommands point to it, and `-h` works like `--help`). `feynman model set` now surfaces actionable diagnostics when a model is unusable — missing provider credentials (pointing to the `feynman model login` command), a local provider in `models.json` lacking an `apiKey` placeholder, or close-match suggestions for an unknown model ID, e.g. `gpt-5.6` suggests `openai/gpt-5.6-terra`.

Discover which models are available when you don't know the exact model ID to pass.

```
$ feynman model
```

List all model subcommands and their usage after upgrading to v0.5.15.

```
$ feynman model help
```

Set a specific provider/model as the default for all Feynman research sessions without being prompted.

```
$ feynman model set openai/gpt-4o
```

Discover available models interactively before committing to a default — useful when you don't know the exact model ID.

```
$ feynman model set
```

Browse and select a model interactively when you don't know the exact model ID.

```
$ feynman model set
```

See all model subcommands when you're unsure what's available under `feynman model`.

```
$ feynman model help
```

— Exact commands, flags and error cases all named. v0.5.15

02 **npm reinstall requirement for Pi dependency** BREAKING 75

npm installs with `legacy-peer-deps` enabled will skip installing Pi; affected users must reinstall Feynman with `--legacy-peer-deps=false` to resolve the resulting missing package.

— Exact flag and fix given, migration-ready. v0.5.15

03 **MCP server support via Pi upgrade** NEW 60

Feynman adds MCP server support through the `/mcp` endpoint, delivered by upgrading to Pi 0.99.1. The Pi dependency is now pinned so future Pi releases only reach Feynman after passing its own checks.

— Endpoint and version named but no usage example. v0.5.15

04 **Sequential OpenAlex search rate limiting** IMPROVED 60

Parallel OpenAlex semantic searches now run sequentially at a one-per-second rate limit and retry once on a 429, fixing a bug where five of six searches would fail.

— Clear mechanism and before/after, but nothing for reader to invoke. v0.5.15

└──▷ BREAKING ON UPGRADE

! npm installs that have `legacy-peer-deps` enabled will skip Pi; `feynman` now requires reinstalling with `--legacy-peer-deps=false` to resolve the missing package.

v0.5.15 NOTES ↗ LATEST STABLE

Feynman v0.5.15 adds MCP server support, improves `feynman model set` with a picker and diagnostic messages, and serializes OpenAlex searches to fix 429 failures.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.5.15 https://github.com/Companion-Inc/feynman.git
# already have the repo? check out this version:
$ git checkout v0.5.15
```

└──▷ **TRY IT**

Browse and select a model interactively when you don't know the exact model ID.

```
$ feynman model set
```

See all model subcommands when you're unsure what's available under `feynman model`.

```
$ feynman model help
```

Adds `feynman model set` with no argument opens an interactive picker of available models instead of failing.

Adds `feynman model help` to list all model subcommands; unknown model subcommands point to it, and `-h` works like `--help`.

Adds MCP server support via Pi 0.99.1's `/mcp` endpoint, now declared and pinned in Feynman's dependency.

`feynman model set` now explains why a model is unavailable — missing provider credentials (with the `feynman model login` command to run), a `models.json` local provider missing an `apiKey`, or an unknown ID with close-match suggestions (e.g. `gpt-5.6` suggests `openai/gpt-5.6-terra`).

Parallel OpenAlex semantic searches now run sequentially at the one-per-second rate limit and retry once on a 429, so all searches succeed instead of five of six failing.

└──▷ **BREAKING ON UPGRADE**

! npm installs that have `legacy-peer-deps` enabled will skip Pi; `feynman` now requires reinstalling with `--legacy-peer-deps=false` to resolve the missing package.

WAS THIS USEFUL? 👍 👎

Databricks Omnigent

Feedback ⓘ v0.16.0 NOTES ↗ RANK

Open-source meta-harness that orchestrates Claude Code, Codex, Cursor, Pi and custom agents behind one interface, with policy controls, OS-level sandboxing and live session sharing.

> Omnigent added GitHub Copilot as a supported coding-agent harness and copy-on-write sandbox edits, alongside new Slack and CORS environment variables and Pi agent prompt controls, while two breaking changes restrict MCP server management to session owners and cap saved preferences and inference snapshots at 65,535 bytes.

└──▷ **WHAT SHIPPED · 13 FEATURES** most completely described first ? what's the number?

01 **GitHub Copilot as a coding-agent harness** NEW 80

Adds the `omni copilot` shorthand command, equivalent to `omni run --harness copilot`, to launch GitHub Copilot as a coding agent through Omnigent's harness layer. Requires installing the `copilot` extra package.

— Names exact command, flag and required package product docs

02 **Pi agent prompt and context controls** NEW 80

Adds `executor.context_files: false` to disable auto-discovered context files and `executor.system_prompt_mode: replace` to replace the base system prompt entirely for Pi agent configurations.

Override the default Pi agent system prompt entirely and silence auto-discovered context files for a tightly scoped security-review agent.

yaml

```yaml
executor:
  system_prompt_mode: replace
  context_files: false
```

— Two config keys with a runnable YAML example v0.16.0

03 **Slack default agent and host environment variables** NEW 75

Adds `OMNIGENT_SLACK_DEFAULT_AGENT_ID` and `OMNIGENT_SLACK_DEFAULT_HOST_TYPE` environment variables to pre-select the Slack setup agent and managed host for new workspaces.

Pre-wire a self-hosted Slack integration so every new workspace automatically uses the correct agent and host type.

```env
OMNIGENT_SLACK_DEFAULT_AGENT_ID=ag-prod-helpdesk
OMNIGENT_SLACK_DEFAULT_HOST_TYPE=managed
```

— Named env vars with a runnable configuration example v0.16.0

04 **Wildcard WebSocket CORS origins** IMPROVED 75

Adds `*.` wildcard entries in `OMNIGENT_WS_ALLOWED_ORIGINS` to trust entire subdomains for WebSocket CORS instead of listing each origin individually.

Allow all subdomains of your company domain as trusted WebSocket origins without listing each one individually.

```env
OMNIGENT_WS_ALLOWED_ORIGINS=*.corp.example.com
```

— Named env var with wildcard syntax example v0.16.0

05 **Copy-on-write sandbox edits** NEW 70

Adds `copy_on_write: true` under `write_paths` so Linux sandbox agents can make disposable edits without mutating the underlying filesystem.

— Config key and behavior named, no runnable example v0.16.0

06 **Saved preferences and snapshots size cap** BREAKING 65

Saved preferences and inference snapshots are now capped at 65,535 bytes: on upgrade, oversized preferences are reset to defaults and oversized snapshots are cleared after compression, and new oversized project-order saves or snapshots are rejected before session creation.

— Exact byte limit and upgrade behavior, no command v0.16.0

THINNER COVERAGE BELOW

07 **MCP server management restricted to session owners** BREAKING 55

Only session owners can now manage MCP servers or replace agent bundles; collaborators who previously could edit MCP configuration now have read-only visibility and must ask the owner to make changes.

— Clear before/after permission change, no migration steps v0.16.0

08 **Workspace browser and file panel redesign** IMPROVED

50

Adds a unified workspace browser and compact file panel with badges for added, modified, and deleted files, plus a Show in Finder action for local files.

— Names specific UI badges and Finder action v0.16.0

09 **Admin default session visibility control** NEW 50

Adds an admin control to default new sessions to public visibility (off / sandbox / all), configured from Settings > Sharing.

— Names UI path and three visibility options v0.16.0

10 **Desktop onboarding redesign** IMPROVED 45

Redesigns desktop onboarding with returning-user detection, in-app CLI install on macOS, organization-preset servers, and a Join your team path.

— Lists onboarding elements but no exact steps v0.16.0

11 **Shallow and blobless sandbox clone options** NEW 35

Adds shallow and blobless repository clone options for admins in sandbox configuration.

— No config key or flag named, thin description v0.16.0

12 **Native image support for MCP tools and embedded browser** NEW 30

Adds MCP tool images and embedded-browser screenshots as native images to supported harnesses.

— States the addition but no mechanism or config v0.16.0

13 **Opt-in in-app browser for chat links** NEW 25

Adds an opt-in setting to open chat links in the in-app browser on desktop.

— Bare description of a toggle, no location given v0.16.0

└──▷ **BREAKING ON UPGRADE**

! Saved preferences and inference snapshots are now capped at 65,535 bytes: on upgrade, oversized preferences are reset to defaults, oversized snapshots are cleared after compression, and new oversized project-order saves or snapshots are rejected before session creation.

! Only session owners can manage MCP servers or replace agent bundles; collaborators who previously could edit MCP configuration now have read-only visibility and must ask the owner to make changes.

v0.16.0 NOTES ↗ LATEST STABLE

Omnigent v0.16.0 adds sandbox copy-on-write edits, wildcard CORS origins, Slack env-var presets, and richer Pi agent prompt controls.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.16.0 https://github.com/omnigent-ai/omnigent.git
# already have the repo? check out this version:
$ git checkout v0.16.0
```

└──▷ **USE IT**

Allow all subdomains of your company domain as trusted WebSocket origins without listing each one individually.

.env

```
OMNIGENT_WS_ALLOWED_ORIGINS=*.corp.example.com
```

Pre-wire a self-hosted Slack integration so every new workspace automatically uses the correct agent and host type.

.env

```
OMNIGENT_SLACK_DEFAULT_AGENT_ID=ag-prod-helpdesk
OMNIGENT_SLACK_DEFAULT_HOST_TYPE=managed
```

Override the default Pi agent system prompt entirely and silence auto-discovered context files for a tightly scoped security-review agent.

yaml

```
executor:
  system_prompt_mode: replace
  context_files: false
```

Adds `copy_on_write: true` in `write_paths` so Linux sandbox agents make disposable edits without mutating the underlying filesystem.

Adds `*.` wildcard entries in `OMNIGENT_WS_ALLOWED_ORIGINS` to trust entire subdomains for WebSocket CORS.

Adds `OMNIGENT_SLACK_DEFAULT_AGENT_ID` and `OMNIGENT_SLACK_DEFAULT_HOST_TYPE` environment variables to pre-select the Slack setup agent and managed host.

Adds `executor.context_files: false` to Pi agent config to disable auto-discovered context files.

Adds `executor.system_prompt_mode: replace` to Pi agent config to replace the base system prompt entirely.

+6 more

└─▷ **BREAKING ON UPGRADE**

! Saved preferences and inference snapshots are now capped at 65,535 bytes: on upgrade, oversized preferences are reset to defaults, oversized snapshots are cleared after compression, and new oversized project-order saves or snapshots are rejected before session creation.

! Only session owners can manage MCP servers or replace agent bundles; collaborators who previously could edit MCP configuration now have read-only visibility and must ask the owner to make changes.

WAS THIS USEFUL? 👍 👎

deepset Haystack

Feedback ⓘ v3.3.0-rc1 NOTES ↗ RANK

Haystack builds language-model applications with modular pipelines for retrieval-augmented generation and agents.

> Haystack's v3.3.0-rc1 release cuts query overhead in SentenceWindowRetriever and tightens BM25L/BM25Plus matching semantics, the latter being a breaking change for existing thresholds.

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

01 BM25L/BM25Plus retrieval now excludes non-matching documents

75

BREAKING

`InMemoryDocumentStore.bm25_retrieval` and `InMemoryBM25Retriever` with `BM25L` (the default) or `BM25Plus` now return only documents containing at least one query term, matching the original BM25L/BM25+ definitions. Result sets may be smaller than `top_k`, and matching-document scores are lower than before, so fixed score thresholds should be re-checked; `BM25Okapi` is unaffected.

– Names classes, methods and migration note but no numeric example v3.3.0-rc1

02 SentenceWindowRetriever batched Document Store queries

IMPROVED 60

`SentenceWindowRetriever` now queries the Document Store once per `run` or `run_async` call instead of once per retrieved document, reducing latency and Document Store load with no code changes required.

– Names the mechanism and method but no benchmark numbers v3.3.0-rc1

└──▷ **BREAKING ON UPGRADE**

! `InMemoryDocumentStore.bm25_retrieval` and `InMemoryBM25Retriever` with `BM25L` (the default) or `BM25Plus` now return only documents containing at least one query term; result sets may be smaller than `top_k`, and matching-document scores are lower than before. If you filter results with a fixed score threshold, re-check the threshold. `BM25Okapi` is not affected.

> v3.3.0-rc1 NOTES ↗ PRE-RELEASE
>
> SentenceWindowRetriever cuts Document Store queries from one-per-document to one-per-run, lowering latency for existing pipelines.
>
> └──▷ **GET THIS VERSION**

```
$ git clone --branch v3.3.0-rc1 https://github.com/deepset-ai/haystack.git
# already have the repo? check out this version:
$ git checkout v3.3.0-rc1
```

Speeds up `SentenceWindowRetriever` by querying the Document Store once per `run` or `run_async` call instead of once per retrieved document, reducing latency and Document Store load with no code changes required.

`InMemoryDocumentStore.bm25_retrieval` and `InMemoryBM25Retriever` with `BM25L` or `BM25Plus` now return only documents containing at least one query term, matching the original BM25L/BM25+ definitions.

> └──▷ **BREAKING ON UPGRADE**
>
> ! `InMemoryDocumentStore.bm25_retrieval` and `InMemoryBM25Retriever` with `BM25L` (the default) or `BM25Plus` now return only documents containing at least one query term; result sets may be smaller than `top_k`, and matching-document scores are lower than before. If you filter results with a fixed score threshold, re-check the threshold. `BM25Okapi` is not affected.

WAS THIS USEFUL? 👍 👎

Stacklok Mecatl

Feedback ⓘ v0.0.43 NOTES ↗ RANK

Mecatl is an open-source agent harness for running the same provider-agnostic agent loop locally, remotely, or on Kubernetes, with durable state, permissions, and gRPC/HTTP-SSE APIs.

> Mecatl v0.0.43 adds production persistence adapter implementations to support durable session state in production deployments.

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

01 **Production persistence adapters for session state** NEW

28

Mecatl publishes production persistence implementations via the adapters layer, enabling durable session state to be maintained in production deployments.

— No adapter names, config keys, or interfaces specified v0.0.43

v0.0.43 NOTES ↗ LATEST STABLE

Mecatl v0.0.43 publishes production persistence adapter implementations for durable session state.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.0.43 https://github.com/stacklok/mecatl.git
# already have the repo? check out this version:
$ git checkout v0.0.43
```

WAS THIS USEFUL? 👍 👎

Mastra

Feedback ⓘ @mastra/core@1.72.0 NOTES ↗ RANK

Mastra is a TypeScript framework for AI applications and agents, providing workflows, agent memory, RAG, tool calling, and evals.

> Mastra 1.72.0 focuses on operational resilience for agents and workflows — live channel resolution without redeploy, lease-fenced background tasks, a cybersecurity-refusal retry handler, and deeper trace aggregation/filtering — alongside a new Microsoft Teams channel and several breaking API changes.

└──▷ **WHAT SHIPPED · 14 FEATURES** most completely described first ? what's the number?

01 **Trace aggregation endpoint for observability** NEW 90

Adds ObservabilityStorage.aggregateTraces() and a `trace-aggregate` capability declaration, plus a `POST /observability/traces/aggregate` endpoint, with planTraceAggregate() and parseTraceAggregateRequest() helpers for grouped/time-bucketed trace measures.

— Names endpoint method/path and helper functions with clear mechanism

@mastra/core@1.72.0

02 **Lease-fenced background tasks** NEW 85

Adds a `leaseDurationMs` option to Mastra({ backgroundTasks }) and `expectedOwnerId` / `expectedStatus` write conditions to storage.updateTask(), providing persisted, expiring ownership leases that prevent multiple workers from double-running the same background task.

Enable lease-fenced background tasks so multiple workers share one storage without double-running the same task.

typescript

```typescript
const mastra = new Mastra({
  backgroundTasks: { enabled: true, leaseDurationMs: 30_000 },
});
```

— Names config key, storage method and shows a runnable config example

@mastra/core@1.72.0

03 **CyberRefusalHandler retry processor** NEW 80

Introduces the `CyberRefusalHandler` processor, importable from `@mastra/core/processors`, which retries once when OpenAI (`cyber_policy`) or Anthropic (`cyber` classifier, `content-filter` finish reason) cybersecurity safeguards refuse a step during an agent run.

Register CyberRefusalHandler on a coding agent so OpenAI and Anthropic cybersecurity false-positive refusals are retried automatically instead of failing the run.

```typescript
import { Agent } from '@mastra/core/agent';
import { CyberRefusalHandler } from '@mastra/core/processors';

const agent = new Agent({
  id: 'coder',
  name: 'Coder',
  instructions: 'You are a coding agent',
  model: 'openai/gpt-5.6-sol',
  errorProcessors: [new CyberRefusalHandler()],
  outputProcessors: [new CyberRefusalHandler()],
  maxProcessorRetries: 3,
});
```

— Named import path, trigger conditions and runnable agent example `@mastra/core@1.72.0`

04 **Live channel resolvers without redeploy** NEW 75

Mastra now supports a channels() resolver function passed to the Mastra({ channels }) constructor, plus a new mastra.resolveChannels() method, so channel connections added or removed on Mastra Platform appear on a running server without a redeploy.

— Names the resolver and method but no usage example `@mastra/core@1.72.0`

05 **Advanced trace query filters** IMPROVED 75

Adds `runId`, `sessionId`, `userId`, and `organizationId` filter fields to advanced trace queries, usable at trace scope or inside `spans.some` / `spans.none`, with `eq`, `ne`, `in`, `notIn`, `exists`, and `notExists` operators.

— Enumerates fields and operators but no runnable query example `@mastra/core@1.72.0`

06 **Initial history streaming for threads** NEW 75

Adds a `withInitialHistory` option to subscribeToThread() so subscriptions emit a `thread-history` chunk first and skip replaying completed runs on retained backends such as Redis Streams.

Stream thread history to a chat UI before live chunks arrive, avoiding stale approval replays after a server restart.

```typescript
const subscription = await agent.subscribeToThread({
  threadId,
  resourceId,
  withInitialHistory: { perPage: 40 },
});

for await (const chunk of subscription.stream) {
  if (chunk.type === 'thread-history')
renderHistory(chunk.payload.messages);
  else applyLiveChunk(chunk);
}
```

— Named option with worked streaming example `@mastra/core@1.72.0`

07 **Teams channel and multi-connection support in @mastra/connect** NEW 65

Introduces the `@mastra/teams` package (Microsoft Teams channel provider) with 25 Microsoft Teams tools added to `@mastra/connect`, plus multi-connection support in `@mastra/connect` via `connection_name` connection selection, bundling Slack/Telegram/Discord channel dependencies.

— Names package, tool count and selector field, no usage example `@mastra/core@1.72.0`

08 **Failure-tolerant observational memory** NEW 60

Adds `maxRetries` and `failurePolicy` options to Memory observational-memory `observation` and `reflection` settings; setting `failurePolicy: 'continue'` prevents an Observer or Reflector model failure from ending the agent turn.

— Names config options and a concrete value with no code sample `@mastra/core@1.72.0`

THINNER COVERAGE BELOW

09 **Workspace sandbox and tool description additions** NEW 55

Adds a `requireDescription` option to the workspace `execute_command` tool, which when enabled inserts a plain-language `description` field before `command` in the tool schema, and adds an optional `sandboxId` field to the `WorkspaceSandbox` interface so hosts can read and persist a provider's physical, reattachable sandbox ID.

— Names both option and field but describes them only briefly `@mastra/core@1.72.0`

10 **Detachable run observation** NEW 50

Adds a detach() method to DurableAgent.observe()'s return value so an observer can stop watching a run without affecting it; early exit from the `fullStream` `for await`

loop now detaches automatically.

– Names the method and auto-detach behaviour, no example `@mastra/core@1.72.0`

11 **Tool approval now requires call ID** BREAKING 50

session.respondToToolApproval(...) now requires `toolCallId`; approval responses without an ID are rejected.

– Names the required field and rejection behaviour, no migration example

`@mastra/core@1.72.0`

12 **Lightweight workflow run summaries** IMPROVED 45

Adds a `summary` flag to listWorkflowRuns() so storage adapters may return only `{ status, timestamp }` per run snapshot instead of the full snapshot.

– Names the flag and return shape but no example `@mastra/core@1.72.0`

13 **Playground UI component API changes** BREAKING 45

`@mastra/playground-ui` removed or renamed multiple UI APIs: `ScrollArea` props, `PageHeader` props, the `ToolCall` component (renamed to `Activity`), and color token API changes.

– Names affected components but no migration path given `@mastra/core@1.72.0`

14 **Retained topic trimming for Redis Streams** NEW 40

Adds PubSub.trimTopic(), implemented in `@mastra/redis-streams`, to delete finished run entries and prevent retained thread topics from growing unbounded.

– Names the method but no example of invocation `@mastra/core@1.72.0`

└──▷ **BREAKING ON UPGRADE**

! session.respondToToolApproval(...) now requires `toolCallId`; approval responses without an ID are rejected.

! `@mastra/playground-ui` removed or renamed multiple UI APIs: `ScrollArea` props, `PageHeader` props, the `ToolCall` component (renamed to `Activity`), and color token API changes.

`@mastra/core@1.72.0` NOTES ↗ LATEST STABLE

Mastra 1.72.0 adds live channel resolvers, lease-fenced background tasks, CyberRefusalHandler, trace aggregation, Teams channel, and richer thread streaming.

└──▷ **GET THIS VERSION**

```
$ git clone --branch @mastra/core@1.72.0 https://github.com/mastra-ai/mastra.git
# already have the repo? check out this version:
$ git checkout @mastra/core@1.72.0
```

└──▷ **USE IT**

Register CyberRefusalHandler on a coding agent so OpenAI and Anthropic cybersecurity false-positive refusals are retried automatically instead of failing the run.

typescript

```
import { Agent } from '@mastra/core/agent';
import { CyberRefusalHandler } from '@mastra/core/processors';

const agent = new Agent({
  id: 'coder',
  name: 'Coder',
  instructions: 'You are a coding agent',
  model: 'openai/gpt-5.6-sol',
  errorProcessors: [new CyberRefusalHandler()],
  outputProcessors: [new CyberRefusalHandler()],
  maxProcessorRetries: 3,
});
```

Enable lease-fenced background tasks so multiple workers share one storage without double-running the same task.

typescript

```
const mastra = new Mastra({
  backgroundTasks: { enabled: true, leaseDurationMs: 30_000 },
});
```

Stream thread history to a chat UI before live chunks arrive, avoiding stale approval replays after a server restart.

```typescript
const subscription = await agent.subscribeToThread({
  threadId,
  resourceId,
  withInitialHistory: { perPage: 40 },
});

for await (const chunk of subscription.stream) {
  if (chunk.type === 'thread-history')
renderHistory(chunk.payload.messages);
  else applyLiveChunk(chunk);
}
```

Adds channels() resolver function support to the Mastra({ channels }) constructor and a new mastra.resolveChannels() method so channel connections added or removed on Mastra Platform appear on a running server without redeploy.

Adds `leaseDurationMs` option to Mastra({ backgroundTasks }) and `expectedOwnerId` / `expectedStatus` write conditions to storage.updateTask() for persisted, expiring ownership leases that prevent multiple workers from double-running the same background task.

Adds `summary` flag to listWorkflowRuns() so storage adapters may return only `{ status, timestamp }` per run snapshot instead of the full snapshot.

Adds ObservabilityStorage.aggregateTraces() method and `trace-aggregate` capability declaration, plus `POST /observability/traces/aggregate` endpoint, with planTraceAggregate() and parseTraceAggregateRequest() helpers for grouped/time-bucketed trace measures.

Adds `runId`, `sessionId`, `userId`, and `organizationId` filter fields to advanced trace queries, usable at trace scope or inside `spans.some` / `spans.none` with `eq`, `ne`, `in`, `notIn`, `exists`, and `notExists` operators.

+9 more

└─▷ BREAKING ON UPGRADE

- ! session.respondToToolApproval(...) now requires `toolCallId`; approval responses without an ID are rejected.
- ! `@mastra/playground-ui` removed or renamed multiple UI APIs: `ScrollArea` props, `PageHeader` props, the `ToolCall` component (renamed to `Activity`), and color token API changes.

WAS THIS USEFUL? 👍 👎

◆ MCP Servers & Tooling

gram

Feedback ⓘ server@2.14.0 NOTES ↗ RANK

Gram provides MCP servers and AI-agent tools from APIs with access controls and observability.

> Gram's server, dashboard and admin releases this window centered on a new direct-grant-over-role-block access model and a wave of Platform MCP tools covering risk findings, tool usage, skill/plugin distribution and network ingress, alongside MCP 2026-07-28 protocol support and remote MCP server identity client management.

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

01 **Remote MCP server identity client management** NEW 90

Adds `remoteSessions.count` to return how many distinct people hold a live session through one remote session client, surfaced in the dashboard's User Identity row. Adds an optional `registration_method` parameter to `remoteSessions.commitServerIdentityConfiguration` in auto client mode, taking `cimd` (default, prefers Client ID Metadata Document, falls back to dynamic client registration) or `dcr` (always registers dynamically). The dashboard's remote MCP server User Identity client panel was redesigned to show connected state, signed-in count, and scopes, with Auto-Configure support for choosing between CIMD and DCR when the provider supports both.

Force dynamic client registration when configuring a remote MCP server identity client, bypassing CIMD even if the provider supports it.

```
$ curl -X POST https://<your-gram-instance>/api \  -H 'Content-Type: application/json' \  -d '{"mutation": "mutation { remoteSessions { commitServerIdentityConfiguration(registration_method: \"dcr\") { success } } }"}'
```

— Names both API fields, both config values, and includes a runnable example.

dashboard@0.128.0 server@2.14.0

02 **Direct-grant access model and grant visibility** NEW 70

Introduces a direct-grant-over-role-block access model: a grant made directly to a person for a specific resource now outranks a block inherited from a role or from everyone, across all blockable permissions (organization, projects, MCP servers,

environments, skills, plugins, and workloads). Adds `direct_selectors` to `access.listGrants` API responses so each scope's directly assigned grants are visible alongside inherited ones.

– Names the API field and permission scopes; no UI walkthrough given. `dashboard@0.128.0` `server@2.14.0`

---

03 **Risk findings and Watchdog Platform MCP tools** NEW

70

Adds `list_risk_findings`, `list_risk_findings_by_chat`, and `get_risk_rule_breakdown` Platform MCP tools to query the same store as the dashboard's Risk Events listing, returning redacted matched values and pseudonymized user identities. Adds `mark_risk_findings_false_positive` and `unmark_risk_findings_false_positive` so an administrator can dismiss or restore specific Watchdog findings from an agent, backed by an idempotency key, per-finding audit, and a partitioned receipt.

– Five tool names given with purpose, but no invocation example. `server@2.14.0`

---

04 **Tool usage summary Platform MCP tool** NEW

65

Adds `get_tool_usage_summary` Platform MCP tool, breaking one project's tool calls down by target type (hosted MCP servers, tunneled MCP servers, gateways, shadow MCP servers, local tools, and skills) over a named window.

– Names the tool and its breakdown dimensions, no example call. `server@2.14.0`

---

05 **Skill distribution and plugin discovery Platform MCP tools** NEW

60

Adds `list_skill_distributions` and `undistribute_skill` Platform MCP tools for administrators to see which plugins carry which skills and retract a skill from a specific plugin or assistant. Adds `list_plugins`, `get_plugin`, and `list_plugin_assignments` to the project's managed assistant, enabling plugin resolution before skill distribution.

– Five tool names listed but each described in one line only. `server@2.14.0`

---

06 **Existing MCP server discovery and Tailscale migration** IMPROVED

60

Extends the Platform MCP `add-existing-mcp-servers` workflow to work from any agent client, report local (stdio or localhost) servers, and optionally migrate servers onto the organization's Tailscale private network.

– Describes scope and mechanism of the extension, no exact invocation. `server@2.14.0`

THINNER COVERAGE BELOW

07 **Network ingress visibility Platform MCP tool** NEW 55

Adds a read-only `get_network_ingress` Platform MCP tool for organization administrators, reporting private networking state, ingress configured/enabled/observed state, and the next action required for private MCP access.

— Names the tool and its report fields, no usage example. `server@2.14.0`

08 **Staff Admin MCP write consent** NEW 55

Adds guarded staff Admin MCP write consent for off-by-default operations, with same-staff browser approval, immutable proposals, idempotent receipts, and staff audit masking. Extends this to guarded staff Admin MCP writes for `consent_tool_filtering`, onboarding task visibility, and enterprise trial extension, with readable approval previews.

— Approval flow mechanism described but this is an internal staff-only surface.

`server@2.14.0`

09 **Directory role mapping visibility** NEW 55

Adds `directory_role_ids` field on member records to report roles assigned via directory role mapping, separate from directly assigned roles. Shows directory-role-mapped members on the Team page, Roles & Permissions page, role filters, and role member counts, with a directory source indicator.

— Names the field and the exact UI pages that surface it. `dashboard@0.128.0`

10 **Tunneled MCP server identity assertions** IMPROVED 55

Exposes `organization_id` on tunneled MCP server settings and uses the server's saved resource identifier (or `tunneled-mcp-server:<ID>` when empty) as the audience of signed caller assertions.

— Names the field and fallback identifier format precisely. `dashboard@0.128.0`

11 **MCP 2026-07-28 protocol support** IMPROVED 50

Remote MCP URL verification now recognizes servers speaking MCP `2026-07-28`, probing first with `server/discover` and falling back to the `initialize` handshake.

— Names the protocol version and handshake methods but is automatic, not reader-actionable.

`server@2.14.0`

12 **Configurable Slack assistant reply behavior** NEW 50

Adds configurable Slack assistant reply behavior: reply only when @`-mentioned`, when @`-mentioned` and then for the rest of that thread, or to any message in its channels, configurable from chat or trigger settings.

— Names the three modes and where to configure them. `server@2.14.0`

13 **Admin sidebar shortcuts for WorkOS and demo org** NEW

50

Adds `WORKOS_ENVIRONMENT_ID` environment variable support to the admin API, enabling an `Open in WorkOS` link beside the WorkOS org ID on the admin organization record when the org is linked to WorkOS. Adds a Demo organization link to the admin sidebar that opens the explore demo org in a new tab.

— Names the env var and exact sidebar link, but each addition is minor. admin@0.6.1

---

14 **Empty tool list normalization in risk policy** IMPROVED

40

Normalizes an empty tool list in an MCP-scoped risk policy to mean 'every tool on this server' instead of rejecting the policy or silently dropping the server from scope.

— Explains the before/after behavior but is a narrow fix. dashboard@0.128.0

---

15 **Access member lookup Platform MCP tool** NEW

30

Admits the Platform MCP `list_access_members` tool to the project's managed assistant for looking up organization members by an identity query of at least three characters.

— Single thin bullet with minimal mechanism. server@2.14.0

---

16 **Hook fail-open default for new organizations** IMPROVED

30

Enables hook fail-open during control-plane outages by default for newly created organizations.

— States the behavior change but no scope beyond one sentence. dashboard@0.128.0

server@2.14.0 NOTES ↗ STABLE

gram server 2.14.0 adds MCP 2026-07-28 support, new Platform MCP tools, and a direct-grant-beats-block access model.

└──▷ GET THIS VERSION

```
$ git clone --branch server@2.14.0 https://github.com/speakeasy-api/gram.git
# already have the repo? check out this version:
$ git checkout server@2.14.0
```

Adds `remoteSessions.count` to return how many distinct people hold a live session through one remote session client, surfaced in the dashboard's User Identity row.

Adds `remoteSessions.commitServerIdentityConfiguration` optional `registration_method` parameter in auto client mode: `cimd` (default, prefers Client ID Metadata Document, falls back to dynamic client registration) or `dcr` (always registers dynamically).

Adds `access.listGrants` reporting of each scope's `direct_selectors`, enabling visibility into per-person access rules.

Adds `list_risk_findings`, `list_risk_findings_by_chat`, and `get_risk_rule_breakdown` Platform MCP tools to query the same store as the dashboard's Risk Events listing, with redacted matched values and pseudonymized user identities.

Adds `mark_risk_findings_false_positive` and `unmark_risk_findings_false_positive` Platform MCP tools so an administrator can dismiss or restore specific Watchdog findings from an agent, with idempotency key, per-finding audit, and a partitioned receipt.

+11 more

dashboard@0.128.0 NOTES ↗ STABLE

gram: Adds `direct_selectors` to `access.listGrants` responses so each scope's directly assigned grants are visible alongside inherited ones.

└──▷ **GET THIS VERSION**

```
$ git clone --branch dashboard@0.128.0 https://github.com/speakeasy-api/gram.git
# already have the repo? check out this version:
$ git checkout dashboard@0.128.0
```

└──▷ **TRY IT**

Force dynamic client registration when configuring a remote MCP server identity client, bypassing CIMD even if the provider supports it.

```
$ curl -X POST https://<your-gram-instance>/api \  -H 'Content-Type: application/json' \  -d '{"mutation": "mutation { remoteSessions { commitServerIdentityConfiguration(registration_method: \"dcr\") { success } } }"}'
```

Adds `directory_role_ids` field on member records to report roles assigned via directory role mapping, separate from directly assigned roles.

Redesigns the remote MCP server User Identity client panel to show connected state, signed-in count, and scopes, and adds Auto-Configure support for choosing

between CIMD and DCR when the provider supports both.

Shows directory-role-mapped members on the Team page, Roles & Permissions page, role filters, and role member counts, with a directory source indicator.

Enables hook fail-open during control-plane outages by default for newly created organizations.

Exposes `organization_id` on tunneled MCP server settings and uses the server's saved resource identifier (or `tunneled-mcp-server:<ID>` when empty) as the audience of signed caller assertions.

+1 more

admin@0.6.1 NOTES ↗ STABLE

Admin sidebar gains a Demo org shortcut and a direct WorkOS org link when `WORKOS_ENVIRONMENT_ID` is configured.

└──▷ **GET THIS VERSION**

```
$ git clone --branch admin@0.6.1 https://github.com/speakeasy-api/gram.git
# already have the repo? check out this version:
$ git checkout admin@0.6.1
```

Adds `WORKOS_ENVIRONMENT_ID` environment variable support to the admin API, enabling an `Open in WorkOS` link beside the WorkOS org ID on the admin organization record when the org is linked to WorkOS.

Adds a Demo organization link to the admin sidebar that opens the explore demo org in a new tab.

WAS THIS USEFUL? 👍 👎

Composio

Feedback ⓘ `@composio/cli@0.4.3-beta.409` NOTES ↗ RANK

Composio connects AI agents to external tools and services with managed authentication and integrations.

> Composio's biggest changes this window were a Tool Router session upgrade across the Google and Cloudflare providers and a reworked instant-charge/account model in core, alongside a sweep of deprecated OpenAI Assistants helpers, 25 new toolkit integrations, and SDK telemetry improvements.

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

01 instantCharge exposed and account metadata renamed

BREAKING

80

`@composio/core@0.22.0` exposes `instantCharge` on execution results, exposes the exact Instant charge as `instantCharge` in project usage summaries and as `metadata.instant_charge` in tool log metadata, and renames search account metadata to `instantAccount` (use `instant_account` for explicit account selection), aligned with `instant.return_instant_charge` on session creation and updates. Code referencing the old field name will break.

— Names every field and config key but lacks a code sample `@composio/core@0.22.0`

02 25 new toolkit integrations NEW

60

Composio's toolkit gains 25 new integrations: `Anakin`, `BeatAPI`, `Browser Use`, `Cohere`, `Coinbase Agents MCP`, `Copper`, `Creem`, `Financial Datasets AI MCP`, `FuseAI`, `FuseAI MCP`, `Harness`, `Higgsfield`, `Jobber`, `Kintone REST`, `Mihu AI`, `MongoDB MCP`, `Ngram`, `OPNForm`, `Statsig`, `Temporal`, `vidIQ MCP`, `Voiceflow`, `Wispr Flow MCP`, `withPICA`, and `WorkOS`.

— Names all 25 integrations but no usage detail per one `product docs`

THINNER COVERAGE BELOW

03 SDK client and runtime telemetry identification NEW

20

The Composio SDK now identifies client products and runtime dimensions for telemetry and observability purposes.

— Only a bare description, no mechanism or surface named `@composio/cli@0.4.3-beta.409`

└──▷ **BREAKING ON UPGRADE**

! Custom provider subclasses overriding `executeToolCall` (Google/Cloudflare) or `handleResponse` (OpenAI Responses) now receive session targets as arguments and may require updates to their method signatures.

! Session-passing behaviour requires `@composio/core` 0.17.0 or later; upgrading without also upgrading `@composio/core` will not work.

! Custom provider subclasses overriding `executeToolCall` or `handleResponse` may require updates because these methods now accept session targets — and passing a session requires `@composio/core` 0.17.0 or later.

! The `handleAssistantMessage`, `waitAndHandleAssistantToolCalls`, and `waitAndHandleAssistantStreamToolCalls` helpers are removed from `OpenAIProvider`; callers must switch to `OpenAIResponsesProvider` from `@composio/openai`.

! Search account metadata is renamed from the previous name to `instantAccount`; code referencing the old field name will break.

`@composio/cli@0.4.3-beta.409` NOTES ↗ PRE-RELEASE

Composio SDK now identifies client products and runtime dimensions for telemetry.

└──▷ **GET THIS VERSION**

```
$ git clone --branch @composio/cli@0.4.3-beta.409
https://github.com/ComposioHQ/composio.git
# already have the repo? check out this version:
$ git checkout @composio/cli@0.4.3-beta.409
```

Adds client product and runtime dimension identification to the SDK for telemetry and observability.

`@composio/google@0.12.0` NOTES ↗ LATEST STABLE

Google and Cloudflare providers' `executeToolCall` now accepts a Tool Router session, aligning with other provider helpers.

└──▷ **GET THIS VERSION**

```
$ git clone --branch @composio/google@0.12.0
https://github.com/ComposioHQ/composio.git
# already have the repo? check out this version:
$ git checkout @composio/google@0.12.0
```

Enables `executeToolCall` on the Google and Cloudflare providers to accept a Tool Router session, so session meta-tools retain their session context through these helpers.

Makes the Cloudflare provider's `options` argument optional in `executeToolCall`.

└─▷ **BREAKING ON UPGRADE**

- ! Custom provider subclasses overriding `executeToolCall` (Google/Cloudflare) or `handleResponse` (OpenAI Responses) now receive session targets as arguments and may require updates to their method signatures.
- ! Session-passing behaviour requires `@composio/core` 0.17.0 or later; upgrading without also upgrading `@composio/core` will not work.

`@composio/core@0.22.0` NOTES ↗ STABLE

Composio @composio/core@0.22.0 exposes `instantCharge` on execution results and usage summaries, and drops deprecated OpenAI Assistants helpers.

└─▷ **GET THIS VERSION**

```
$ git clone --branch @composio/core@0.22.0
https://github.com/ComposioHQ/composio.git
# already have the repo? check out this version:
$ git checkout @composio/core@0.22.0
```

Exposes `instantCharge` on execution results and renames search account metadata to `instantAccount`; use `instant_account` for explicit account selection, aligned with `instant.return_instant_charge` on session creation and updates.

Exposes the exact Instant charge as `instantCharge` in project usage summaries and as `metadata.instant_charge` in tool log metadata.

Removes deprecated `OpenAIProvider` helpers `handleAssistantMessage`, `waitAndHandleAssistantToolCalls`, and `waitAndHandleAssistantStreamToolCalls`; migrate to `OpenAIResponsesProvider` from `@composio/openai` with the Responses API.

└─▷ **BREAKING ON UPGRADE**

- ! The `handleAssistantMessage`, `waitAndHandleAssistantToolCalls`, and `waitAndHandleAssistantStreamToolCalls` helpers are removed from `OpenAIProvider`; callers must switch to `OpenAIResponsesProvider` from `@composio/openai`.
- ! Search account metadata is renamed from the previous name to `instantAccount`; code referencing the old field name will break.

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ Agentic Infrastructure

Perplexity API

Feedback ⓘ snapshot-20260930 NOTES ↗ RANK

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

> Perplexity's Agent API gained support for two new models, `openai/gpt-6.1-sol` and `anthropic/claude-sonnet-5-5`, usable via the chat completions endpoint.

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

01 **New Agent API model support** NEW 80

The Agent API now supports `openai/gpt-6.1-sol` and `anthropic/claude-sonnet-5-5` as selectable models. Both can be invoked via the `https://api.perplexity.ai/chat/completions` endpoint by setting the `model` field to `openai/gpt-6.1-sol` or `anthropic/claude-sonnet-5-5` in the request body.

Query GPT-6.1 Sol via the Perplexity API chat completions endpoint.

```
$ curl https://api.perplexity.ai/chat/completions -H
'Authorization: Bearer <your_api_key>' -H 'Content-Type:
application/json' -d '{"model": "openai/gpt-6.1-sol", "messages":
[{"role": "user", "content": "Summarize the OWASP Top 10."}]}'
```

Query Claude Sonnet 5.5 via the Perplexity API chat completions endpoint.

```
$ curl https://api.perplexity.ai/chat/completions -H
'Authorization: Bearer <your_api_key>' -H 'Content-Type:
application/json' -d '{"model": "anthropic/claude-sonnet-5-5",
"messages": [{"role": "user", "content": "List common
misconfigurations in AWS S3 buckets."}]}'
```

— Names exact model IDs and endpoint with runnable examples. snapshot-20260930

snapshot-20260930 NOTES ↗ LATEST STABLE

Perplexity Agent API adds support for `openai/gpt-6.1-sol` and `anthropic/claude-sonnet-5-5` models.

Adds `openai/gpt-6.1-sol` as a supported model in the Agent API.

> Adds `anthropic/claude-sonnet-5-5` as a supported model in the Agent API.

WAS THIS USEFUL?  

Together AI

Feedback ⓘ snapshot-20260930 NOTES ↗ RANK

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

> Together AI raised the maximum LoRA fine-tuning rank to 128 with new inspection tooling, let the CLI deploy fine-tuned models by registry name, and introduced spend controls for account billing.

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

01 **LoRA rank limit raised to 128** BREAKING 95

Together AI raises the maximum LoRA adapter rank from 64 to 128 for the majority of models; set `lora_r` to use a rank above the default of 64. The model limits response now includes a `lora_training.default_rank` field alongside `lora_training.max_rank`, and the new `tg fine-tuning model-limits <model>` CLI command inspects a model's default and maximum LoRA rank. CLI and Python SDK versions earlier than 2.36.0 (including the 1.x SDK) use the model's maximum rank when `lora_r` is not set, so any workflow that omits `lora_r` will now silently train at rank 128 instead of 64 — upgrade to SDK/CLI 2.36.0 or explicitly set `lora_r`, especially when continuing training from a rank-64 adapter.

Check a model's default and maximum LoRA rank before launching a fine-tuning job.

```
$ tg fine-tuning model-limits <model>
```

— Names exact fields, CLI command, and breaking-change mechanism. product docs

THINNER COVERAGE BELOW

02 **CLI deploy by fine-tuning registry name** IMPROVED 45

The Together CLI now accepts a completed fine-tuning job's qualified registry name in place of its model ID when deploying a fine-tuned model.

— Names the capability but no command example given. product docs

03 **Spend controls for account billing** NEW 30

Together AI adds spend controls supporting balance cut-offs, auto-recharge, cost attribution, and programmatic monitoring to manage Together AI spend.

— Lists capabilities but no config keys, endpoints, or usage details. snapshot-20260930

└──▷ **BREAKING ON UPGRADE**

! CLI and Python SDK versions earlier than 2.36.0 (including the 1.x SDK) use the model's maximum rank when `lora_r` is not set; now that maximum is 128 (up from 64), any existing workflow that omits `lora_r` will silently train at rank 128 instead of 64. Upgrade to SDK/CLI 2.36.0 or explicitly set `lora_r`, especially when continuing training from a rank-64 adapter.

> `snapshot-20260930` NOTES ↗ LATEST STABLE
>
> Together AI adds spend controls with balance cut-offs, auto-recharge, cost attribution, and programmatic monitoring.

WAS THIS USEFUL?  

FlashAttention

Feedback ⓘ fa4-v4.0.0.beta33 NOTES ↗ RANK

FlashAttention implements fast, memory-efficient exact attention kernels used by most current transformer training and inference stacks.

> FlashAttention's fa4-v4.0.0.beta33 release focuses on SM100 (Blackwell) MLA kernel coverage, adding 1CTA MLA forward with a new dispatch heuristic and extending sparse MLA backward support, alongside a Hopper forward-tiling throughput optimization.

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

01 **Sparse MLA backward coverage on SM100** IMPROVED 68

Extends sparse MLA backward to work at any head count on SM100, expands Sparse MLA recompute-P to allow 64 Q heads on SM100, and adds `seqused_q` / `seqused_k` support in SM100 HD256 backward.

– Names concrete parameters and scope limits lifted, no usage example fa4-v4.0.0.beta33

THINNER COVERAGE BELOW

02 **1CTA MLA forward and dispatch heuristic on SM100** NEW 48

Adds 1CTA MLA forward supporting both dense and sparse top-k attention on SM100, along with a new MLA dispatch heuristic to select between kernel variants.

– Names kernel variants but no flag or config to invoke directly fa4-v4.0.0.beta33

03 **Hopper forward tile fitted to block-sparse block size** IMPROVED 45

On SM90 (Hopper), fits the forward tile to the block-sparse block size, skipping the causal/local mask on unmasked KV blocks for better throughput.

– Describes mechanism and benefit but no user-facing control fa4-v4.0.0.beta33

fa4-v4.0.0.beta33 NOTES ↗ PRE-RELEASE

FlashAttention fa4-v4.0.0.beta33 adds 1CTA MLA forward (dense + sparse top-k), MLA dispatch heuristic, and sparse MLA backward at any head count on SM100.

└──▷ **GET THIS VERSION**

```
$ git clone --branch fa4-v4.0.0.beta33 https://github.com/Dao-AILab/flash-attention.git
# already have the repo? check out this version:
$ git checkout fa4-v4.0.0.beta33
```

Adds `seqused_q` / `seqused_k` support in SM100 HD256 backward.

Expands Sparse MLA recompute-P to allow 64 Q heads on SM100.

Fits forward tile to block-sparse block size on SM90 (Hopper), skipping the causal/local mask on unmasked KV blocks for better throughput.

WAS THIS USEFUL? 👍 👎

Eigen Labs Darkbloom

Feedback ⓘ v0.9.12 NOTES ↗

CODE ↗ RANK

Private Inference Network on Idle Macs

> Darkbloom's v0.9.12 release introduces cached-prompt billing with new pricing fields, native MiMo V2.6 Flash support, and a round of coordinator scheduling and diagnostics improvements.

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

01 **Cached-prompt billing for prefix cache hits** NEW 85

Prompt tokens served from the prefix cache now bill at a per-model `cache_read_price` instead of the full input rate, surfaced to consumers as `prompt_tokens_details.cached_tokens` and advertised to OpenRouter as `pricing.input_cache_read` in the provider feed.

— Names exact config key and API fields, no numeric limits given v0.9.12

02 **Provider model load readiness in status and doctor commands**

NEW 75

Explains provider model load readiness (required memory, usable memory, and shortfall) in `darkbloom status`, `darkbloom doctor`, and the My Macs view.

— Names exact CLI commands and UI path, immediately runnable v0.9.12

THINNER COVERAGE BELOW

03 **Coordinator routing by first-content forecasts and work measurements**

IMPROVED 50

Routes coordinator requests by first-content forecasts, using verified prompt work and observed rates for admission decisions, and adds provider work measurements and measured warm placement to improve request scheduling.

— Explains mechanism but no config surface or metric names given v0.9.12

04 **Cache routing state persists across coordinator restarts**

IMPROVED 50

Persists the cache routing holder and demand indexes across coordinator restarts so the holder index is not empty after a restart.

— Describes fix mechanism but no config key or command given v0.9.12

05 **Default SSD cache write budget raised to 750 GB/day** IMPROVED

45

Raises the default SSD cache write budget to 750 GB/day.

— Gives exact number but no config key or override method v0.9.12

06 **Automated model weight revisions with staged rollback** NEW

35

Automates model weight revisions with staged activation and rollback.

— States mechanism briefly, no config or command named v0.9.12

07 **Native support for MiMo V2.6 Flash** NEW

30

Adds native provider support and prompt contracts for MiMo V2.6 Flash.

— Names model but gives no mechanism or usage detail v0.9.12

08 **Unbounded-projection reasons preserved in request profiles**

IMPROVED

25

Preserves engine unbounded-projection reasons in request profiles.

— Bare description, no mechanism or usage given v0.9.12

v0.9.12 NOTES ↗ CODE ↗ LATEST STABLE

Darkbloom v0.9.12 adds cached-prompt billing, MiMo V2.6 Flash support, and first-content forecast routing

└──▷ GET THIS VERSION

```
$ git clone --branch v0.9.12 https://github.com/Layr-Labs/d-inference.git
# already have the repo? check out this version:
$ git checkout v0.9.12
```

Adds cached prompt pricing: prompt tokens served from the prefix cache now bill at a per-model `cache_read_price` instead of the full input rate, surfaced to consumers as `prompt_tokens_details.cached_tokens` and advertised to OpenRouter as `pricing.input_cache_read` in the provider feed.

Adds native provider support and prompt contracts for MiMo V2.6 Flash.

Routes coordinator requests by first-content forecasts, using verified prompt work and observed rates for admission decisions.

Adds provider work measurements and measured warm placement to improve request scheduling.

> Automates model weight revisions with staged activation and rollback.
>
> +4 more

WAS THIS USEFUL? 👍 👎

ArcBox

Feedback ⓘ v0.8.0 NOTES ↗ RANK

ArcBox runs containers, microVM agent sandboxes, and full Linux or macOS virtual machines on a Mac from one Rust daemon and CLI.

> ArcBox v0.8.0 adds native SSH access with vsock agent forwarding, local CA-backed TLS termination for container domains, a Firecracker VM driver backend, and k3s LoadBalancer support, alongside NFS rootfs export, new Docker config options, and a string of internal networking, snapshot, and sandbox-lifecycle improvements.

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

01 **Native SSH access to VMs and containers** NEW 91

ArcBox now runs an SSH server authenticated with the daemon's key, supporting shell, exec, sftp, `pty-req`, `window-change`, `env`, `direct-tcpip` port forwarding, and signal delivery. `abctl ssh install` adds ArcBox's SSH config to the OpenSSH client config, and ssh-agent is forwarded into containers over a vsock slot pool that relays the container's `ssh-auth` socket.

Install the ArcBox SSH config so you can `ssh <machine>@arcbox` directly from your terminal using your existing OpenSSH client.

```
$ abctl ssh install
```

— Names full SSH feature set and exact command, but no limits given v0.8.0

02 **Docker daemon configuration in config.toml** NEW 81

The `[docker]` section of `~/.config/arcbox/config.toml` gains an `expose_ports_to_lan` key so operators can bind published ports to `127.0.0.1` instead of every interface, and a new `[docker.engine]` section lets operators configure the guest `dockerd` with any `daemon.json` key.

Keep Docker published ports bound to loopback only so they are never reachable from the local network — useful on shared or untrusted Wi-Fi.

```toml
[docker]
expose_ports_to_lan = false
```

— Exact config keys and file path with a runnable example v0.8.0

03 **TLS termination and routing for container domains** NEW

75

ArcBox generates a name-constrained local CA before the System VM boots and uses it to terminate TLS for container domains; `abctl tls trust` and `abctl tls untrust` manage that CA in the system trust store. It also now picks a container domain's HTTP port from the container's actual listeners and routes port 80 to it.

— Named commands and CA mechanism, no UI walkthrough given v0.8.0

---

04 **Sandbox lifecycle and resource metering** IMPROVED 73

Adds a `storage_bytes` field to meter the retained disk footprint of a sandbox in every lifecycle state, adds a monotonic sequence number to `SandboxEvent` in the computer protocol, enables runtime changes to the System VM's CPU and memory limits without a restart (persisting the new size to `config.toml`), and makes handover a distinct lifecycle transition in the computer subsystem.

— Names four concrete fields/behaviors, no trigger command v0.8.0

---

05 **Namespace debug shell for containers** NEW 70

`abctl debug` opens a shell inside any running container by entering its namespaces directly, letting operators inspect a container's filesystem, processes, or network without installing tooling into the image.

Drop into a live shell inside any running container to inspect its filesystem, processes, or network without installing tooling into the image.

```
$ abctl debug <container>
```

— Exact command given and mechanism named, little else v0.8.0

THINNER COVERAGE BELOW

---

06 **k3s LoadBalancer service support** NEW 59

ArcBox supports k3s `servicelb` to run LoadBalancer Services and forwards Kubernetes LoadBalancer ports to the host; the agent RPC protocol gains a LoadBalancer Service listing.

— Names servicelb and RPC addition, no command shown v0.8.0

---

07 **Firecracker VM driver backend** NEW 50

Introduces the Firecracker VM driver adapter (`fc-driver`, R1 part 1) as a new backend for the VM driver port, adding an alternative virtualization backend.

— Names the adapter but gives no usage detail v0.8.0

---

08 **VM network relaying and egress policy** IMPROVED 50

ArcBox now relays TCP connections from inside machines as flow-controlled sessions and forwards `direct-tcpip` channels to ports inside machines, and makes guest egress proxying an explicit network policy rather than always-on behavior.

— Explains behavior change but no config key to control it v0.8.0

09 **NFS export of container rootfs** NEW 45

Each running container's rootfs is exported over NFS with NFS-exportable overlay mount options, making container filesystems readable from the host.

— Mechanism described but no command or path given v0.8.0

10 **New backend adapters for tap-net and snapshot subsystems** NEW 45

Adds an `nftables` PacketFilter backend to the tap-net subsystem and a `util-linux` BlockTools backend to the snapshot subsystem.

— Named backends but no behavioral detail given v0.8.0

11 **Computer subsystem adapter and exception cleanup** BREAKING 33

The runtime stops naming an adapter, and the last two exceptions in the computer subsystem are retired; existing configurations relying on those adapter names or exception paths will break on upgrade.

— Breaking change noted but no specific names given v0.8.0

└──▷ BREAKING ON UPGRADE

! The runtime stops naming an adapter, and the last two exceptions in the computer subsystem are retired — existing configurations relying on those adapter names or exception paths will break on upgrade.

v0.8.0 NOTES ↗ LATEST STABLE

ArcBox v0.8.0 adds native SSH access, TLS termination for container domains, a local CA, container namespace debug shells, NFS rootfs export, Firecracker VM driver, and k3s L

└──▷ GET THIS VERSION

```
$ git clone --branch v0.8.0 https://github.com/arcboxlabs/arcbox.git
# already have the repo? check out this version:
$ git checkout v0.8.0
```

└──▷ TRY IT

Drop into a live shell inside any running container to inspect its filesystem, processes, or network without installing tooling into the image.

```
$ abctl debug <container>
```

Keep Docker published ports bound to loopback only so they are never reachable from the local network — useful on shared or untrusted Wi-Fi.

~/.config/arcbox/config.toml

```
[docker]
expose_ports_to_lan = false
```

Install the ArcBox SSH config so you can `ssh <machine>@arcbox` directly from your terminal using your existing OpenSSH client.

```
$ abctl ssh install
```

Adds `abctl debug` to open a shell inside any container by entering its namespaces directly.

Adds `abctl ssh install` to include ArcBox's SSH config in the OpenSSH client config.

Adds `abctl tls trust` and `abctl tls untrust` subcommands to manage the local CA in the system trust store.

Adds `expose_ports_to_lan` key to `~/.config/arcbox/config.toml` under `[docker]` so operators can keep published ports off the LAN (binding to `127.0.0.1` instead of every interface).

Adds `[docker.engine]` section to `~/.config/arcbox/config.toml` to let operators configure the guest `dockerd` with any `daemon.json` key.

+17 more

└─▷ **BREAKING ON UPGRADE**

! The runtime stops naming an adapter, and the last two exceptions in the computer subsystem are retired — existing configurations relying on those adapter names or exception paths will break on upgrade.

WAS THIS USEFUL? 👍 👎

HeyGen HyperFrames

Feedback ⓘ v0.8.97 NOTES ↗ RANK

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

> HyperFrames shipped a set of host-configurable controls for its embedded Studio timeline and canvas, alongside smaller timeline UI additions like a drop-target lane and pointed popovers.

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

01 **Host customization controls for embedded Studio** NEW

65

Hosts embedding Studio can now restrict the timeline tool menu to only the Select and Split tools (hiding the rest), lock the timeline to read-only while keeping seek, scroll, and zoom interactive, turn off body drags while keeping resize, rotate, and crop handles active, hide timeline history, position the shortcuts button, and set track padding.

— Describes several host options but no named flags or config keys. v0.8.97 v0.8.94

THINNER COVERAGE BELOW

02 **Drop-target lane in Studio timeline** NEW 50

Adds an empty drop-target lane rendered below the last track in the Studio timeline, with a tightened top gap, so media can be dragged directly onto it.

— Clear UI behavior described but no API surface named. v0.8.95

03 **Directional popover arrows** NEW 35

Adds an optional arrow to popovers so they can visually point at their trigger element.

— Simple one-line UI addition with minimal detail. v0.8.97

v0.8.97 NOTES ↗ LATEST STABLE

HyperFrames v0.8.97 adds rightward clip selection with host-trimmed tool menus, popover arrows, and a read-only timeline mode.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.8.97 https://github.com/heygen-
com/hyperframes.git
# already have the repo? check out this version:
$ git checkout v0.8.97
```

Enables hosts to restrict the timeline tool menu to only the Select and Split tools, hiding the rest from end users.

Adds an optional arrow to popovers so they can visually point at their trigger element.

Lets hosts lock the timeline to read-only while keeping seek, scroll, and zoom interactive.

v0.8.95 NOTES ↗ STABLE

HyperFrames v0.8.95 adds a visible drop-target lane below the last Studio timeline track.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.8.95 https://github.com/heygen-
com/hyperframes.git
# already have the repo? check out this version:
$ git checkout v0.8.95
```

v0.8.94 NOTES ↗ STABLE

HyperFrames v0.8.94 adds host controls to disable body drags and customize timeline appearance in embedded Studio.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.8.94 https://github.com/heygen-
com/hyperframes.git
# already have the repo? check out this version:
$ git checkout v0.8.94
```

Enables a host to turn off body drags while keeping resize, rotate, and crop handles in embedded Studio.

Lets a host hide timeline history, position the shortcuts button, and set track padding in embedded Studio.

WAS THIS USEFUL? 👍 👎

OmniRoute

Feedback ⓘ v3.8.51 NOTES ↗ RANK

OmniRoute puts hundreds of model providers behind one endpoint, with quota-aware fallback and prompt compression.

> OmniRoute significantly deepened its routing intelligence and provider reach this window: the catalog grew to 357 providers, new auto/* fallback modes and a modelVisibilityAllowlist/Denylist control plane shipped, Video Bridge gained a drilldown API plus audio extraction and bounded transcript provenance, and a native macOS menu-bar app and a batch of CLI commands for launching and configuring coding agents arrived alongside dozens of smaller integrations.

└──▷ **WHAT SHIPPED · 32 FEATURES** most completely described first ? what's the number?

01 **Video Bridge drilldown API** NEW 95

Adds authenticated `POST /api/v1/video-bridge/drilldown` consumer route with opaque hashed handles, preview/standard/detail multi-resolution variants, response pagination capped at 8 frames and 32 MiB, and cross-key access denial.

— Full endpoint, limits and access model specified v3.8.51

02 **CLI commands for launching and configuring coding agents** NEW 90

`omniroute run` launches any of 7 coding CLIs (Claude Code, Codex, Aider, Goose, OpenCode, Qwen Code, Gemini CLI) in one command; `omniroute configure` adds an interactive provider+model picker supporting 10 targets and per-context favorites; `omniroute models` queries the remote server's model registry; 13 registered `setup-*` commands cover CLI/agent onboarding; and a per-tool setup reference for all 36 tools (26 CLI Code + 10 CLI Agents) is published at `docs/reference/CLI-TOOLS.md`.

Launch Claude Code through OmniRoute so all traffic is routed and quota-tracked without touching your global node install.

```
$ omniroute run
```

Fan out a single prompt to five providers simultaneously with `auto/chaos` to compare outputs or maximise availability.

```
$ omniroute configure codex
```

— Names exact commands, counts and doc path product docs

03 **New auto/* routing modes and route prefixes** NEW 90

`auto/chaos` fans a prompt out in parallel to a panel of models (one per provider, 5 by default) and returns a single answer; `auto/lkgp` pins to the last successful provider then falls back to routing rules; `auto/subscription` routes only through plan-included connections with a hard-stop on overage; `auto/thrifty` steps up `subscription → keyless → free → cheap → premium` as each tier is exhausted; an `expiry-first` account fallback strategy prioritises routing; and the `oc/…` (OpenCode Free) route prefix calls a specific free backend directly.

Route all traffic through plan-included connections first, stepping up only when each tier is exhausted, to minimise surprise overage charges.

json

```
{
  "model": "auto/thrifty",
  "messages": [{"role": "user", "content": "Summarise this
report."}]
}
```

— Names every routing mode and its exact fallback order v3.8.51

04 **modelVisibilityAllowlist and modelVisibilityDenylist settings**

NEW 90

Adds `modelVisibilityAllowlist` and `modelVisibilityDenylist` settings to curate exactly which models `GET /v1/models` advertises, mirrored into every `auto/*` combo candidate pool so a denied model cannot be routed via combo selection.

Lock down which models your tenant exposes and can route to — useful for compliance environments that must never touch unapproved models.

json

```
{
  "modelVisibilityAllowlist": ["gpt-4o", "claude-3-5-sonnet"],
  "modelVisibilityDenylist": ["gpt-4-32k"]
}
```

— Exact config keys, endpoint and enforcement scope given v3.8.51

05 **OmniRouteTray macOS menu-bar app** NEW 85

OmniRouteTray is a macOS menu-bar app, installable via `brew install --cask zoispag/tap/omniroute-tray`, that supervises `omniroute serve`, auto-updates in place with rollback, and shows live provider quota bars plus Claude session/weekly limits.

Install OmniRouteTray on macOS so the gateway starts on login and stays alive without a terminal window.

```
$ brew install --cask zoispag/tap/omniroute-tray
```

Install the macOS menu-bar tray so OmniRoute starts at login and stays alive without a terminal window.


```
$ brew install --cask zoispag/tap/omniroute-tray
```


— Exact install command and supervised behaviour given `product docs`

06 **SeekAi OpenAI-compatible gateway integration** NEW 80

Adds SeekAi (`seekai.cc`) as an OpenAI-compatible New-API gateway with catalog id `seekai` (alias `ska`) at `https://seekai.cc/v1` , using live `passthroughModels` via `/v1/models` .

— Exact endpoint, catalog id and alias given `v3.8.51`

07 **Bounded transcript provenance for Video Bridge** NEW 80

Adds `embedded` transcript provenance earned via a bounded, allowlisted (`mov_text` / `subrip` / `webvtt`) subtitle probe through the loopback-only Video Bridge broker, normalized through a ReDoS-safe WebVTT parser and Zod-validated end to end. Enforces a bounded deterministic contract for Video Bridge transcripts: 256 cues, 4096 input code units and 4 KiB UTF-8 per cue, 64 KiB total text, with malformed-Unicode rejection and a structural provenance trust boundary so caller JSON cannot self-assert `embedded` / `audio-bridge` provenance.

— Exact limits and validation boundary spelled out `v3.8.51`

08 **Free-tier dashboard and free-provider rankings** IMPROVED 75

The `/dashboard/free-tiers` page shows free-tier token budgets — initially ~1.62B tokens/month computed across 17 pools with published monthly budgets plus five per-model Groq caps, then expanded to 489 free-tier entries across 35 recurring pool keys. A separate Free Provider Rankings page adds LMArena-ELO rankings for free providers and now shows actual usage served over the last 24 hours alongside ELO rank, with a dash shown when the sample is too small.

— Exact page path and numbers across three releases `v3.8.51`

09 **GLM-5.3-Flash Coding Plan support in zai** NEW 75

Adds GLM-5.3-Flash Coding Plan support in `zai` (1M context, 128K output, vision, `low|high|max` reasoning) routed through the OpenAI-compatible Coding Plan endpoint.

— Exact model specs and reasoning levels named v3.8.51

---

10 **Video Bridge audio extraction and Audio Bridge STT** NEW

75

Adds opt-in dual-gated Video Bridge audio extraction and Audio Bridge STT: a loopback-only broker `mode=audio` operation extracts a bounded mono 16 kHz PCM WAV from the already-downloaded video, then reuses the existing Audio Bridge transcription boundary.

— Mechanism and audio format specified, opt-in gating noted v3.8.51

---

11 **Build tooling: Turbopack bundler and React Compiler** IMPROVED

75

Adds `OMNIROUTE_USE_TURBOPACK` environment variable to control Turbopack vs. Webpack bundler selection in production builds, exposed as a `--build-arg` in `Dockerfile.bun`; allows the Turbopack bundler flag on Bun 1.4+ with configurable Webpack fallback; and adds `reactCompiler: true` with `babel-plugin-react-compiler` for automatic React memoization at build time.

— Names env var, build-arg and compiler flag v3.8.51

---

12 **16-factor live scoring engine for routing** NEW

70

Adds a 16-factor live scoring engine across every connection covering health, quota, cost, latency, task fit, quality, and session availability (see `docs/routing/AUTO-COMBO.md`).

— Named factors and doc reference, no exact config product docs

---

13 **Custom model and dashboard sort settings** NEW

70

Adds `customModels[].isFree` field to mark custom models as free-tier eligible, and a `config.modelSort` setting (values: manual, provider, score, or name) to control combo model ordering in the dashboard builder, persisted and re-applied on load and after add.

— Named fields and values, no endpoint for direct use v3.8.51

14 **Bounded worker-thread request compression** IMPROVED

70

Runs synchronous RTK and Caveman request compression in a bounded worker-thread pool, keeping large `/v1/responses` compression heaps outside the HTTP isolate with per-engine telemetry.

— Endpoint and mechanism named, no config to invoke v3.8.51

15 **Provider catalog expanded to 357 providers** IMPROVED

65

The provider catalog grew to 357 registered providers (up from 290, with 152 free-tier), adding cloud agent integrations (Codex Cloud, Cursor, Devin, Jules, and Grok Build via xAI) with browser + OAuth login, Ollama as a first-class card, a Kimi official partnership (Code/Web/Moonshot), Claude Opus 5 and Sonnet 5, Zed, Requesty, SenseNova, Yuanbao, and Agnes AI.

— Names every added provider but no usage mechanism product docs

16 **Gemini batch text-to-speech via /v1/audio/speech** NEW

65

Adds Google AI Studio Gemini batch text-to-speech support through `POST /v1/audio/speech`.

— Named endpoint but minimal further detail v3.8.51

17 **Antigravity-family auto-sync default** IMPROVED

65

Defaults new Antigravity-family connections (agy CLI imports and Antigravity OAuth connects) to model auto-sync so live model discovery lands in the synced catalog and `/v1/models` picks up freshly released upstream models without code changes.

— Named endpoint and connection types, default behaviour clear v3.8.51

18 **Focused video analysis mode** NEW

60

Adds an opt-in focused video analysis mode using a normalized, 500-code-point latest-user hint for task-aware frame captions while preserving cache identity without storing raw task text.

— Mechanism and limit named, no exact config key v3.8.51

19 **Deployment guides for Termux, PWA, and Docker** NEW

60

Adds Termux, PWA, and Docker deployment guides at `docs/guides/TERMUX_GUIDE.md`, `docs/guides/PWA_GUIDE.md`, and `docs/guides/DOCKER_GUIDE.md`.

— Exact doc paths given, content itself undetailed product docs

THINNER COVERAGE BELOW

20 **Multi-modal generation API for image, video, and audio** NEW

55

Adds support for image, video, and audio generation via a single API covering xAI Grok Imagine, Novita AI video, ComfyUI, Magnific, Adobe Firefly, Segmind, and ElevenLabs.

— Lists providers but no endpoint or usage detail product docs

21 **Webhook push for request/quota events** NEW 55

Adds webhook push for request/quota events to Slack, Discord, Telegram, or any URL.

— Named destinations but no endpoint or payload spec product docs

22 **Plugin framework and OpenCode provider plugin** NEW

50

Adds a plugin framework and marketplace, plus an OpenCode plugin `@omniroute/opencode-provider` published to npm.

— Names the npm package but framework is undetailed product docs

23 **Durable managed-lease sessions view** NEW 50

Adds a durable exclusive managed-lease sessions view in the existing Sessions dashboard, keeping leased clients visible across idle gaps.

— Describes behaviour but no exact UI path or API v3.8.51

24 **Xquik X search integration** NEW 45

Adds Xquik X search integration with typed results, credential validation, REST routing, and MCP selection.

— Describes capability but no endpoint or config name v3.8.51

25 **Observability: log export and per-attempt timing** NEW 45

Adds scheduled log export to BigQuery, continuous call-log export from the Dashboard, and opt-in per-attempt upstream timing on proxy logs.

— Names three surfaces but no config keys given v3.8.51

26 **Obsidian vault integration** NEW 45

Adds Obsidian vault integration with 22 MCP tools.

— Exact tool count but no config or command product docs

27 **OpenAI-compatible Batch and Files APIs** NEW 35

Adds OpenAI-compatible Batch and Files APIs for asynchronous request processing and file-based model interactions.

— Bare mention with no endpoint or parameters product docs

28 **ACP agent discovery** NEW 35

Adds ACP agent discovery with 15 built-in agents.

— Names a count but no mechanism or usage product docs

29 **Telegram bot bridge** NEW 25

Adds a Telegram bot bridge.

— Bare one-line mention product docs

30 **Semantic response cache** NEW 20

Adds a semantic response cache.

— One-line mention with no mechanism given product docs

31 **Gamification with leaderboards** NEW 20

Adds gamification with leaderboards.

— Bare one-line mention product docs

32 **In-app version manager** NEW 15

Adds an in-app version manager.

— Bare one-line mention with no detail product docs

> v3.8.51 NOTES ↗ LATEST STABLE
>
> OmniRoute v3.8.51 adds subscription-aware auto-routing, model visibility allowlists, a Video Bridge drilldown API, audio extraction/STT, and 30+ provider and

dashboard improv

└──▷ **GET THIS VERSION**

```
$ git clone --branch v3.8.51
https://github.com/diegosouzapw/OmniRoute.git
# already have the repo? check out this version:
$ git checkout v3.8.51
```

└──▷ **USE IT**

Lock down which models your tenant exposes and can route to — useful for compliance environments that must never touch unapproved models.

```json
{
  "modelVisibilityAllowlist": ["gpt-4o", "claude-3-5-sonnet"],
  "modelVisibilityDenylist": ["gpt-4-32k"]
}
```

Route all traffic through plan-included connections first, stepping up only when each tier is exhausted, to minimise surprise overage charges.

```json
{
  "model": "auto/thrifty",
  "messages": [{"role": "user", "content": "Summarise this
report."}]
}
```

Adds `modelVisibilityAllowlist` and `modelVisibilityDenylist` settings to curate exactly which models `GET /v1/models` advertises, mirrored into every `auto/*` combo candidate pool so a denied model cannot be routed via combo selection.

Adds authenticated `POST /api/v1/video-bridge/drilldown` consumer route with opaque hashed handles, preview/standard/detail multi-resolution variants, response pagination capped at 8 frames and 32 MiB, and cross-key access denial.

Adds `auto/subscription` and `auto/thrifty` routing groups — `auto/subscription` routes only through plan-included connections with a hard-stop overage; `auto/thrifty` steps up `subscription → keyless → free → cheap → premium` as each is exhausted.

Adds `customModels[].isFree` field to mark custom models as free-tier eligible.

> Adds SeekAi (`seekai.cc`) as an OpenAI-compatible New-API gateway with catalog id `seekai` (alias `ska`) at `https://seekai.cc/v1` , using live `passthroughModels` via `/v1/models` .
>
> +17 more

WAS THIS USEFUL? 👍 👎

BerriAI LiteLLM

Feedback ⓘ v1.103.1 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 this window focused on supply-chain security for its Docker images, adding cosign signature verification across releases, plus new support for managing OpenAI and Azure OpenAI code interpreter containers.

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

01 Cosign signature verification for Docker images NEW

88

All LiteLLM Docker images are now signed and verifiable with `cosign verify` against a pinned public key, either via the commit-pinned URL `https://raw.githubusercontent.com/BerriAI/litellm/0112e53046018d726492c814b3644b7d376029d0/cosign.pub` or the release-tag URL `https://raw.githubusercontent.com/BerriAI/litellm/v1.103.1/cosign.pub`. Introduced in v1.102.2 and carried through v1.100.4 and v1.103.1, this lets operators cryptographically verify image integrity before deploying to production, e.g. `cosign verify --key <cosign.pub> ghcr.io/berriai/litellm:v1.103.1`.

Verify the integrity of a pulled LiteLLM Docker image against the cryptographically immutable signing key before deploying to production.

```
$ cosign verify \
  --key
https://raw.githubusercontent.com/BerriAI/litellm/0112e53046018d726492c814b3644b7d376029d0/cosign.pub \
  ghcr.io/berriai/litellm:v1.103.1
```

Verify the integrity of a pulled LiteLLM Docker image before deploying it, using the cryptographically immutable commit-pinned public key.

```
$ cosign verify \
  --key
https://raw.githubusercontent.com/BerriAI/litellm/0112e53046018d726492c814b3644b7d376029d0/cosign.pub \
  ghcr.io/berriai/litellm:v1.102.2
```

Verify the integrity of a pulled LiteLLM Docker image before deploying it, using the cryptographically immutable commit-pinned public key.

```
$ cosign verify \
  --key
https://raw.githubusercontent.com/BerriAI/litellm/0112e53046018d7
26492c814b3644b7d376029d0/cosign.pub \
  ghcr.io/berriai/litellm:v1.100.4
```

— Names exact key URLs and a runnable verify command. v1.103.1 v1.102.2 v1.100.4

THINNER COVERAGE BELOW

02 **Code interpreter container management** NEW 30

LiteLLM adds support for managing OpenAI and Azure OpenAI code interpreter containers (sessions) for executing code in isolated environments.

— Describes the capability but no endpoint or usage details. product docs

v1.103.1 NOTES ↗ STABLE

LiteLLM Docker images now support cosign signature verification against a pinned public key.

└──▷ GET THIS VERSION

```
$ git clone --branch v1.103.1 https://github.com/BerriAI/litellm.git
# already have the repo? check out this version:
$ git checkout v1.103.1
```

└──▷ TRY IT

Verify the integrity of a pulled LiteLLM Docker image against the cryptographically immutable signing key before deploying to production.

```
$ cosign verify \
  --key
https://raw.githubusercontent.com/BerriAI/litellm/0112e53046018d7264
92c814b3644b7d376029d0/cosign.pub \
  ghcr.io/berriai/litellm:v1.103.1
```

WAS THIS USEFUL? 👍 👎

TileLang

Feedback ⓘ v0.1.15 NOTES ↗ RANK

TileLang is a Python-based domain-specific language and compiler for writing optimized GPU and CPU kernels.

> TileLang v0.1.15 adds a full Huawei Ascend 950 NPU backend alongside role-based CUDA warp specialization and unified block-scaled GEMM, while installation docs expose new CMake flags for controlling CUDA, ROCm, and Ascend backend builds.

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

01 **Huawei Ascend 950 NPU backend** NEW 90

Adds the `tilelang.ascend.language` module and `target="ascend"` for end-to-end code generation on Huawei Ascend 950 (`dav-3510`) NPUs, with automatic Cube/Vector scheduling, `T.SimdVF` and `T.SimtVF` regions, explicit UB/L1/L0 storage, tiled copies, MXFP8/MXFP4 block-scaled GEMM, Bisheng compilation, and PyTorch NPU tensor and stream support.

Compile and run a kernel on a Huawei Ascend 950 NPU instead of a CUDA GPU.

python

```
import tilelang
import tilelang.ascend.language as T

@tilelang.jit(target="ascend")
def my_kernel(A, B):
    # kernel body using T.SimdVF / T.SimtVF regions
    with T.SimdVF():
        pass
    ...
```

— Full mechanism and code example but no benchmark numbers v0.1.15

02 **CMake flags for CUDA, ROCm, Ascend backends** NEW 90

Installation docs add `-DUSE_CUDA=ON|OFF` (default ON when a CUDA toolkit is found, except macOS, with custom toolkit path support), `-DUSE_ROCM=ON|OFF` (default ON on Linux, with HIP stubs and vendored headers to build without a full ROCm install), and `-DUSE_ASCEND=ON|OFF` (default ON on Linux, with Ascend stubs and local ACL declarations to build without a CANN install). CMake `-D` arguments and existing cache entries take precedence over environment variables, which only initialize options in a fresh build directory.

Build TileLang targeting only the ROCm backend on a machine that has no CUDA toolkit installed.

```
$ cmake .. -DUSE_CUDA=OFF -DUSE_ROCM=ON -DUSE_ASCEND=OFF
```

Disable all three GPU backends to produce a CPU-only build, useful in CI environments with no accelerator hardware.

```
$ cmake .. -DUSE_CUDA=OFF -DUSE_ROCM=OFF -DUSE_ASCEND=OFF
```

— Named flags, defaults, and runnable build commands given `product docs`

03 **Role-based automatic CUDA warp specialization** NEW

88

Setting the `TL_ENABLE_AUTO_WARP_SPECIALIZATION` environment variable to `"role_based"` automatically assigns TMA loads, MMA computation, TMA stores, and worker operations to specialized warp groups within a CUDA kernel.

Enable role-based warp specialization to automatically partition TMA and MMA work across dedicated warp groups in a CUDA kernel.

```python
import os
os.environ["TL_ENABLE_AUTO_WARP_SPECIALIZATION"] = "role_based"

import tilelang
# compile your existing GEMM or FlashAttention kernel normally;
# the scheduler will assign TMA-load, MMA, and TMA-store warp
groups automatically
kernel = my_gemm.compile(M=4096, N=4096, K=4096)
```

— Named env var, mechanism, and working code example `v0.1.15`

04 **Breaking API removals across kernel APIs** BREAKING

85

The `sync` and `group` parameters are removed from `T.Pipelined`, breaking kernels that use them. The `k_pack` parameter is removed from `T.gemm_sp`; ROCm kernels using T.gemm(k_pack=...) must now import `tilelang.rocm.language`. Kernels that query thread extents during tracing must specify `threads=` explicitly in `T.Kernel`. `T.symbolic` is deprecated (alias still present); new code must use `T.dynamic`. `TILELANG_CACHE_VERIFY_HASH` is removed; binary artifact hash verification is now mandatory and legacy cache formats are rebuilt automatically.

— Names every removed/changed API surface with migration guidance v0.1.15

05 **Unified block-scaled GEMM for CUDA** IMPROVED 68

Adds `T.gemm_blockscaled` with unified semantics and dedicated backend dispatch, improved SM100 instruction selection, and expanded SM120 fragment support including fragment-resident A operands and row-major scale fragments.

— Names API and hardware targets but no usage example v0.1.15

06 **Compile-time Python constructs in kernels** IMPROVED 65

The kernel frontend now supports Python compile-time iteration, `enumerate`, `zip`, comprehensions, and generator expressions, while preserving device-loop semantics for for ... in range(...).

— Names constructs but no example shown v0.1.15

THINNER COVERAGE BELOW

07 **DeepSeek and GLM model-specific kernels** NEW 55

Adds an FP8 sparse MLA forward example for DeepSeek V3.2 on Hopper, and supports the DeepSeek V3.2 Top-K selector, wave64 Hadamard transforms, and GLM-5.3 k-pool on ROCm.

— Names specific models and hardware, no usage steps v0.1.15

08 **Round-to-nearest FMA/FMUL intrinsics** NEW 35

Adds `T.fma` and `T.fmul` round-to-nearest intrinsics for CUDA.

— Bare name, no mechanism or example given v0.1.15

09 **JSON kernel parameter caching** IMPROVED 30

Cached kernel parameters are now stored as JSON instead of cloudpickle.

— States format change only, no further detail v0.1.15

10 **32-bit integer atomic add on Metal** NEW 30

Adds 32-bit integer atomic add support on Metal.

— Bare capability statement, no example v0.1.15

11 **Wall-clock benchmarking and MPS timing helpers** NEW 25

Adds wall-clock benchmarking and MPS-compatible timing helpers.

— One-line description, no API names or example v0.1.15

└──▷ BREAKING ON UPGRADE

! The `sync` and `group` parameters are removed from `T.Pipelined`; kernels using them will fail.

! The `k_pack` parameter is removed from `T.gemm_sp`; ROCm kernels using T.gemm(k_pack=...) must import `tilelang.rocm.language`.

! Kernels that query thread extents during tracing must now specify `threads=` explicitly in `T.Kernel`.

! `T.symbolic` is deprecated (alias still present); new code must use `T.dynamic`.

! `TILELANG_CACHE_VERIFY_HASH` is removed; binary artifact hash verification is now mandatory and legacy cache formats are rebuilt automatically.

v0.1.15 NOTES ↗ LATEST STABLE

TileLang v0.1.15 adds Huawei Ascend 950 NPU backend, role-based CUDA warp specialization, unified block-scaled GEMM, and richer Python frontend syntax.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.1.15 https://github.com/tile-ai/tilelang.git
# already have the repo? check out this version:
$ git checkout v0.1.15
```

└──▷ **USE IT**

Compile and run a kernel on a Huawei Ascend 950 NPU instead of a CUDA GPU.

```python
import tilelang
import tilelang.ascend.language as T

@tilelang.jit(target="ascend")
def my_kernel(A, B):
    # kernel body using T.SimdVF / T.SimtVF regions
    with T.SimdVF():
        pass
    ...
```

Enable role-based warp specialization to automatically partition TMA and MMA work across dedicated warp groups in a CUDA kernel.

```python
import os
os.environ["TL_ENABLE_AUTO_WARP_SPECIALIZATION"] = "role_based"

import tilelang
# compile your existing GEMM or FlashAttention kernel normally;
# the scheduler will assign TMA-load, MMA, and TMA-store warp groups automatically
kernel = my_gemm.compile(M=4096, N=4096, K=4096)
```

Adds `tilelang.ascend.language` module and `target="ascend"` to support Huawei Ascend 950 (`dav-3510`) with end-to-end NPU code generation, automatic Cube/Vector scheduling, `T.SimdVF` and `T.SimtVF` regions, explicit UB/L1/L0 storage, tiled copies, MXFP8/MXFP4 block-scaled GEMM, and Bisheng compilation with PyTorch NPU tensor and stream support.

Enables role-based automatic CUDA warp specialization via the `TL_ENABLE_AUTO_WARP_SPECIALIZATION: "role_based"` environment variable, assigning TMA loads, MMA computation, TMA stores, and worker operations to specialized warp groups.

Adds `T.gemm_blockscaled` with unified semantics, dedicated backend dispatch, improved SM100 instruction selection, and expanded SM120 fragment support including fragment-resident A operands and row-major scale fragments.

Adds `T.fma` and `T.fmul` round-to-nearest intrinsics for CUDA.

Supports Python compile-time iteration, `enumerate` , `zip` , comprehensions, and generator expressions in the kernel frontend while preserving device-loop semantics for for ... in range(...).

+5 more

└──▷ **BREAKING ON UPGRADE**

- ! The `sync` and `group` parameters are removed from `T.Pipelined` ; kernels using them will fail.
- ! The `k_pack` parameter is removed from `T.gemm_sp` ; ROCm kernels using T.gemm(k_pack=...) must import `tilelang.rocm.language` .
- ! Kernels that query thread extents during tracing must now specify `threads=` explicitly in `T.Kernel` .
- ! `T.symbolic` is deprecated (alias still present); new code must use `T.dynamic` .
- ! `TILELANG_CACHE_VERIFY_HASH` is removed; binary artifact hash verification is now mandatory and legacy cache formats are rebuilt automatically.

WAS THIS USEFUL?  

OpenAI Python SDK

Feedback ⓘ v3.22.0 NOTES ↗ RANK

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

> OpenAI Python SDK added computer use support to the beta agents API and introduced the GPT-6.1 Sol model identifier.

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

01 **GPT-6.1 Sol model identifier** NEW 50

Adds the `GPT-6.1 Sol` model identifier to the API, making the new model addressable by name in requests.

— Names exact model identifier, usable directly in requests v3.21.0

02 **Computer use in beta agents API** NEW 30

The beta agents API now supports computer use capability.

— Names the API area but no mechanism or usage detail v3.22.0

v3.22.0 NOTES ↗ LATEST STABLE

OpenAI Python SDK v3.22.0 adds computer use support to the beta agents API.

└──▷ GET THIS VERSION

```
$ git clone --branch v3.22.0 https://github.com/openai/openai-python.git
# already have the repo? check out this version:
$ git checkout v3.22.0
```

v3.21.0 NOTES ↗ STABLE

OpenAI Python SDK v3.21.0 adds the GPT-6.1 Sol model identifier.

└──▷ GET THIS VERSION

```
$ git clone --branch v3.21.0 https://github.com/openai/openai-python.git
# already have the repo? check out this version:
$ git checkout v3.21.0
```

WAS THIS USEFUL?  

OpenAI TypeScript SDK

Feedback ⓘ v7.25.0 NOTES ↗

RANK

The OpenAI TypeScript SDK provides typed JavaScript and TypeScript clients for OpenAI APIs and streaming responses.

> OpenAI's TypeScript SDK v7.24.0–v7.25.0 added computer use support to beta Agents and introduced a Realtime Translation WebSocket API, alongside opt-in incremental response snapshots and smaller additions to Agents and the Responses API.

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

01 Realtime translation WebSocket and client secrets NEW

60

Adds a dedicated Node translation WebSocket (`realtime`) for real-time translation workloads, plus support for creating Realtime Translation client secrets via the API.

— Names the `realtime` WebSocket and client secret creation surface. v7.24.0

THINNER COVERAGE BELOW

02 Incremental response snapshots over WebSocket NEW

55

Adds opt-in WebSocket incremental output snapshots to the `responses` API for streaming use cases.

— Names the opt-in mechanism and `responses` API but no config detail. v7.24.0

03 Computer use support in beta Agents NEW

25

v7.25.0 adds computer use capability to the beta Agents SDK.

— No mechanism, flag, or example given beyond the capability name. v7.25.0

04 Cyber access programs support in Responses API NEW

20

Adds Cyber access programs support to the Responses API.

— Bare mention with no mechanism or usage detail. v7.24.0

05 GPT-6.1 Sol model identifier NEW

20

Adds the GPT-6.1 Sol model identifier.

— Names the model but gives no usage or capability detail. v7.24.0

06 Credential and session options for Agents IMPROVED

15

v7.24.0 adds credential and session configuration options for Agents.

— Only named in the release summary, no specifics given. v7.24.0

v7.25.0 NOTES ↗ LATEST STABLE

OpenAI TypeScript SDK v7.25.0 adds computer use support to beta agents.

└──▷ GET THIS VERSION

```
$ git clone --branch v7.25.0 https://github.com/openai/openai-node.git
# already have the repo? check out this version:
$ git checkout v7.25.0
```

v7.24.0 NOTES ↗ STABLE

v7.24.0 adds Realtime translation WebSocket, incremental response snapshots, credential/session options for Agents, and the GPT-6.1 Sol model identifier.

└──▷ GET THIS VERSION

```
$ git clone --branch v7.24.0 https://github.com/openai/openai-node.git
# already have the repo? check out this version:
$ git checkout v7.24.0
```

Adds a dedicated Node translation WebSocket (`realtime`) for real-time translation workloads.

Adds opt-in WebSocket incremental output snapshots to the `responses` API for streaming use cases.

Adds Realtime Translation client secret creation via the API.

Adds Cyber access programs support to the Responses API.

WAS THIS USEFUL? 👍 👎

AWS Amazon Bedrock

Feedback ⓘ Added GPT-6.1 Sol (2026-09-29)

BLOG ↗ RANK

Amazon Bedrock provides managed foundation-model inference, customization, knowledge bases, and APIs for generative AI applications.

GPT-6.1 Sol on Bedrock includes OpenAI first-party base pricing, a 10% regional and US CRIS premium, and OpenAI-aligned lifecycle terms.

NEW INTEGRATION GPT-6.1 Sol model integration via Mantle endpoint on Amazon Bedrock

WAS THIS USEFUL? 👍 👎

Snowflake Cortex AI

Feedback ⓘ

Sep 29, 2026: Response caching for Cortex AI Functions · NOTES ↗ · RANK

Snowflake Cortex AI provides managed language-model functions, search, and agents that work with data governed in Snowflake.

> Snowflake Cortex AI now automatically caches `AI_COMPLETE` results within a query to cut token costs and latency.

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

01 **Automatic response caching for AI_COMPLETE** NEW 55

`AI_COMPLETE` text calls now have automatic response caching: identical calls within a single query return stored results, reducing token usage, credits billed, and latency, with no configuration required.

— Names the function and behaviour but no config surface or limits

Sep 29, 2026: Response caching for Cort…

> Sep 29, 2026: Response caching for Cortex AI Functions · NOTES ↗ · LATEST
>
> STABLE
>
> Snowflake Cortex now automatically caches `AI_COMPLETE` results within a query, cutting token costs and latency for repeated identical calls.
>
> Adds automatic response caching for `AI_COMPLETE` text calls — identical calls within a single query return stored results, reducing token usage, credits billed, and latency with no configuration required.

WAS THIS USEFUL? 👍 👎

Ollama

Feedback ⓘ v0.35.0 NOTES ↗ CODE ↗ RANK

Local large language model runner that downloads, serves, and chats with models through a CLI and REST API.

> Ollama v0.35.0 introduces a new decision-model API for classification tasks and speeds up the settings page.

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

01 **Decision models via POST /v1/systemone** NEW 95

Adds a `POST /v1/systemone` endpoint (based on TypeSafe's Jev API) accepting `model`, `state`, and `questions` fields, returning structured `answers` with `choice`, `probabilities`, and `confidence`. It supports three question types — `choice`, `noul`, and `score` — and two decision models, `nimble` (Bespoke Labs) and `tev1` (Together AI), both installable via `ollama pull`.

Route a support ticket to the correct label by sending context and a `choice` question to the decision model endpoint.

```
$ curl http://localhost:11434/v1/systemone \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "nimble",
    "state": "Our checkout has returned 500 errors since 9am.",
    "questions": {
      "label": {
        "type": "choice",
        "instructions": "Which label fits this ticket?",
        "criteria": {
          "billing": "Payments and refunds",
          "bug": "Software errors",
          "account": "Login and account access"
        }
      }
    }
  }'
```

Pull the `nimble` decision model before sending classification requests.

```
$ ollama pull nimble
```

— Names endpoint, fields, models, and includes runnable examples. v0.35.0

THINNER COVERAGE BELOW

02 **Faster settings page load** IMPROVED 25

The settings page now opens without waiting for model discovery to complete.

— Brief prose note with no mechanism or metric given. v0.35.0

v0.35.0 NOTES ↗ CODE ↗ STABLE

Ollama v0.35.0 adds decision models via `POST /v1/systemone`, returning choices, probabilities, and scores for classification tasks.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.35.0 https://github.com/ollama/ollama.git
# already have the repo? check out this version:
$ git checkout v0.35.0
```

└──▷ **TRY IT**

Route a support ticket to the correct label by sending context and a `choice` question to the decision model endpoint.

```
$ curl http://localhost:11434/v1/systemone \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "nimble",
    "state": "Our checkout has returned 500 errors since 9am.",
    "questions": {
      "label": {
        "type": "choice",
        "instructions": "Which label fits this ticket?",
        "criteria": {
          "billing": "Payments and refunds",
          "bug": "Software errors",
          "account": "Login and account access"
        }
      }
    }
  }'
```

Pull the `nimble` decision model before sending classification requests.

```
$ ollama pull nimble
```

> Adds `POST /v1/systemone` endpoint (based on TypeSafe's Jev API) accepting `model`, `state`, and `questions` fields, returning structured `answers` with `choice`, `probabilities`, and `confidence` — supports three question types: `choice`, `noul`, and `score`.
>
> Supports two decision models: `nimble` (Bespoke Labs) and `tev1` (Together AI), both available via `ollama pull`.
>
> Settings page now opens without waiting for model discovery to complete.

WAS THIS USEFUL? 👍 👎

Axolotl AI Axolotl

Feedback ⓘ v0.20.0 NOTES ↗ RANK

Axolotl fine-tunes open language models from YAML configuration, supporting full fine-tuning, LoRA, QLoRA, and multi-GPU training recipes.

> Axolotl v0.20.0 ships a large batch of YAML-config-driven training features — GGUF export, Ringmaster context parallelism, native NVFP4 LoRA, backend-agnostic expert parallelism, MixLoRA, Dynamic Fine-Tuning, and several memory/throughput tweaks — alongside a set of breaking changes around minimum versions, FSDP1, ring-attention, and legacy Mamba checkpoints.

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

01 **GGUF export via `axolotl export`** NEW 90

Adds `axolotl export config.yml --quantize Q4_K_M,Q8_0` to convert a trained checkpoint to GGUF for llama.cpp, Ollama, LM Studio, and llamafile, emitting one file per requested quant type.

Export a trained checkpoint to GGUF with multiple quantization levels for use with llama.cpp or Ollama.

```
$ axolotl export config.yml --quantize Q4_K_M,Q8_0
```

— Runnable command with quant formats and target runtimes named. v0.20.0

02 **Context parallelism (Ringmaster) for long-context training** 90

BREAKING

Adds `context_parallel_size` and `context_parallel:` block config keys to enable Ulysses, Ring, or USP context parallelism for long-context packed-sequence training via `pip install axolotl[ringmaster]`, with experimental support for recurrent layers (GDN, KDA, Mamba2). The `ring-flash-attn` extra is replaced by `ringmaster`; `ring_attn_func` and `heads_k_stride` are now ignored with a warning.

— Names config keys, extras, algorithms, and the migration path. v0.20.0

03 **New cloud training providers (Baseten, Nebius)** NEW 85

Adds `axolotl train --cloud` support for Baseten and (experimentally) Nebius Serverless Jobs alongside Modal; new providers can ship as plugins via the `axolotl.cloud_providers` entry point, configured with `provider:`, `image`, or `image_build` in the cloud config. Modal is now moved out of core dependencies,

requiring `pip install axolotl[modal]`, and Baseten users need `pip install axolotl[baseten]`.

— Command, plugin entry point, config keys, and install steps named. v0.20.0

04 **FSDP1 removed** BREAKING 85

FSDP1 is removed: `fsdp_version: 1`, the top-level `fsdp` list, `sharding_strategy` (use `reshard_after_forward`), and `forward_prefetch` now raise; `sync_module_states` and `use_orig_params` are dropped with a warning.

— Every removed and replacement config key is named. v0.20.0

05 **NVFP4 LoRA training support** NEW 80

Adds `nvfp4_merge_aware: false` config key to opt out of merge-aware NVFP4 LoRA training (on by default for supported ModelOpt/SonicMoE and native TorchAO paths) under FSDP2, DeepSpeed ZeRO-1/2/3, or tensor parallelism, and adds LoRA-training support for `nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4` (Nemotron-3 LatentMoE) through ScatterMoE or SonicMoE, with routed experts staying packed NVFP4 and the rest dequantizing to bf16.

— Names the config key, model id, and backends but no command. v0.20.0

06 **Expert parallelism without DeepEP** NEW 80

Adds `expert_parallel_backend: auto` and `expert_parallel_dispatch_chunks` config keys to run expert parallelism via plain `all_to_all` on any GPU setup, composable with FSDP2, HSDP, and context parallelism; expert parallelism now requires `fsdp_version: 2` with `state_dict_type: FULL_STATE_DICT`, and combining it with tensor parallelism is rejected.

— Named config keys and the hard FSDP2 constraint. v0.20.0

07 **Per-dataset subsampling weight field** NEW 80

Adds a `weight` field on SFT dataset entries (value in (0, 1.0]) to subsample a corpus to weight * len(dataset) examples before tokenization without materializing a pre-mixed dataset.

Downsample a large SFT corpus to 20% of its examples before tokenization without pre-mixing datasets.

```yaml
datasets:
  - path: my_org/large_corpus
    type: alpaca
    weight: 0.2
  - path: my_org/small_curated
    type: alpaca
    weight: 1.0
```

— Named field, value range, and a working YAML example. v0.20.0

08 **Legacy Mamba checkpoint loading removed** BREAKING 80

Legacy `MambaLMHeadModel` loading is removed; raw `mamba_ssm` checkpoints such as `state-spaces/mamba-2.8b` with `model_type: MambaLMHeadModel` no longer load — use Hugging Face format checkpoints (e.g. `state-spaces/mamba-130m-hf`) instead.

— Names the removed loader and the exact replacement checkpoint. v0.20.0

09 **TorchAO NF4 backend for FSDP2 QLoRA** IMPROVED 75

Adds `nf4_backend: torchao` config key to use torchao's NF4Tensor as an alternative to bitsandbytes for FSDP2 QLoRA; `qlora_sharded_model_loading` now defaults on for FSDP2 QLoRA with `cpu_ram_efficient_loading`.

— Names the config keys and default change but no run example. v0.20.0

10 **MixLoRA adapter for routed LoRA experts** NEW 75

Adds `adapter: mixlora` with `mixlora_num_experts` and `mixlora_top_k` config keys and the `MixLoraPlugin` to turn each FFN of a dense model into a routed mixture of LoRA experts.

— Named adapter and config keys with a clear mechanism. v0.20.0

11 **FLA backend for Mamba training** NEW 75

Adds `model_config.mamba_backend: fla` config key to swap in FLA mixers for full fine-tuning and LoRA on `in_proj` / `out_proj` for Mamba and Mamba2, with packing and context parallelism support via `pip install axolotl[fla]`.

— Named config key, target modules, and install extra. v0.20.0

12 **fp32 gradient accumulation for large vocabularies** NEW 75

Adds `cut_cross_entropy_accum_c_fp32` and `cut_cross_entropy_c_grad_chunk_size: auto` config keys to accumulate

`lm_head` gradients in fp32 for large vocabularies and bound peak memory.

— Named keys and stated purpose of bounding peak memory. v0.20.0

13 **fp32 LoRA gradients option** NEW 65

Adds `lora_fp32_gradients` config key to keep LoRA parameters and gradients in fp32 across single-device, DDP, FSDP2, and DeepSpeed.

— Named config key and supported backends, thin mechanism. v0.20.0

14 **Python 3.12 and torch 2.13.0 minimums** BREAKING 65

Python 3.12 and torch 2.13.0 are the new minimums; `requires-python` moves from 3.10 to 3.12 and torch from 2.11.0 to 2.13.0.

— Exact version numbers given, no migration steps. v0.20.0

15 **Selective checkpointing activation controls** NEW 60

Adds `save_modules` and `save_matmul_min_k` config keys to control which activations selective checkpointing retains.

— Named keys with only a one-line description of effect. v0.20.0

16 **Sample packing for Mamba model family** NEW 60

Adds sample packing support for Mamba, Mamba2, and Falcon-Mamba Hugging Face checkpoints, with recurrent and convolution state reset at document boundaries.

— Explains the mechanism but no config key named. v0.20.0

17 **Dynamic tokenization work queue** IMPROVED 60

Adds a dynamic work queue for dataset tokenization when `dataset_num_proc > 1`, so workers pull chunks as they finish instead of using fixed shards.

— Explains the before/after scheduling behaviour clearly. v0.20.0

THINNER COVERAGE BELOW

18 **Dynamic Fine-Tuning (DFT) loss reweighting** NEW 55

Adds `use_dft: true` config key for Dynamic Fine-Tuning, reweighting each token's cross-entropy by its own detached probability.

— Config key and mechanism given, no further detail. v0.20.0

19 **Liger kernels and fused attention for GLM-4.7-Flash** NEW

55

Adds Liger RoPE, RMSNorm, and SwiGLU support plus `fused_attn_kernel` for GLM-4.7-Flash.

— Named model and flag, but only a list of kernels. v0.20.0

20 **scao optimizer support** NEW

40

Adds `optimizer: scao` support after `pip install scao`.

— Bare name and install step, no mechanism given. v0.20.0

21 **New example configs for additional models** NEW

40

Adds example configs for MiniCPM5-2B, Nex-N2.5-mini, Qwen3.6-35B-A3B, IFM K2-Horizon (0.9B/3.7B/7B/32B dense and MoVA sparse-MoE), and ERNIE-4.5 0.3B SFT.

— Just a list of model names with example configs. v0.20.0

└──▷ **BREAKING ON UPGRADE**

! Python 3.12 and torch 2.13.0 are the new minimums; `requires-python` moves from 3.10 to 3.12 and torch from 2.11.0 to 2.13.0.

! FSDP1 is removed: `fsdp_version: 1`, the top-level `fsdp` list, `sharding_strategy` (use `reshard_after_forward`), and `forward_prefetch` now raise; `sync_module_states` and `use_orig_params` are dropped with a warning.

! The `ring-flash-attn` extra is replaced by `ringmaster`; `ring_attn_func` and `heads_k_stride` are now ignored with a warning. Install `axolotl[ringmaster]` for context parallelism.

! Legacy `MambaLMHeadModel` loading is removed; raw `mamba_ssm` checkpoints such as `state-spaces/mamba-2.8b` with `model_type: MambaLMHeadModel` no longer load — use Hugging Face format checkpoints (e.g. `state-spaces/mamba-130m-hf`) instead.

! `modal` is moved out of core dependencies; Modal cloud users now need `pip install axolotl[modal]`, and Baseten users need `pip install axolotl[baseten]`.

! Expert parallelism (EP) now requires `fsdp_version: 2` with `state_dict_type: FULL_STATE_DICT`; EP with tensor parallelism is rejected.

v0.20.0 NOTES ↗ LATEST STABLE

Axolotl v0.20.0 adds GGUF export, Ringmaster context parallelism, native NVFP4 LoRA, expert parallelism without DeepEP, and more.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.20.0 https://github.com/axolotl-ai-cloud/axolotl.git
# already have the repo? check out this version:
$ git checkout v0.20.0
```

└──▷ **TRY IT**

Export a trained checkpoint to GGUF with multiple quantization levels for use with llama.cpp or Ollama.

```
$ axolotl export config.yml --quantize Q4_K_M,Q8_0
```

Downsample a large SFT corpus to 20% of its examples before tokenization without pre-mixing datasets.

config.yml

```
datasets:
  - path: my_org/large_corpus
    type: alpaca
    weight: 0.2
  - path: my_org/small_curated
    type: alpaca
    weight: 1.0
```

Adds `axolotl export config.yml --quantize Q4_K_M,Q8_0` to convert a trained checkpoint to GGUF for llama.cpp, Ollama, LM Studio, and llamafile, emitting one file per requested quant type.

Adds `context_parallel_size` and `context_parallel:` block config keys to enable Ulysses, Ring, or USP context parallelism for long-context packed-sequence training via `pip install axolotl[ringmaster]`, with experimental support for recurrent layers (GDN, KDA, Mamba2).

Adds `nvfp4_merge_aware: false` config key to opt out of merge-aware NVFP4 LoRA training (on by default for supported ModelOpt/SonicMoE and native TorchAO paths) under FSDP2, DeepSpeed ZeRO-1/2/3, or tensor parallelism.

Adds `expert_parallel_backend: auto` and `expert_parallel_dispatch_chunks` config keys to run expert parallelism via plain `all_to_all` on any GPU setup, composable with FSDP2, HSDP, and context parallelism.

Adds `nf4_backend: torchao` config key to use torchao's `NF4Tensor` as an alternative to bitsandbytes for FSDP2 QLoRA; `qlora_sharded_model_loading` now defaults on for FSDP2 QLoRA with `cpu_ram_efficient_loading`.

+14 more

└─▷ **BREAKING ON UPGRADE**

- ! Python 3.12 and torch 2.13.0 are the new minimums; `requires-python` moves from 3.10 to 3.12 and torch from 2.11.0 to 2.13.0.
- ! FSDP1 is removed: `fsdp_version: 1`, the top-level `fsdp` list, `sharding_strategy` (use `reshard_after_forward`), and `forward_prefetch` now raise; `sync_module_states` and `use_orig_params` are dropped with a warning.
- ! The `ring-flash-attn` extra is replaced by `ringmaster`; `ring_attn_func` and `heads_k_stride` are now ignored with a warning. Install `axolotl[ringmaster]` for context parallelism.
- ! Legacy `MambaLMHeadModel` loading is removed; raw `mamba_ssm` checkpoints such as `state-spaces/mamba-2.8b` with `model_type: MambaLMHeadModel` no longer load — use Hugging Face format checkpoints (e.g. `state-spaces/mamba-130m-hf`) instead.
- ! `modal` is moved out of core dependencies; Modal cloud users now need `pip install axolotl[modal]`, and Baseten users need `pip install axolotl[baseten]`.
- ! Expert parallelism (EP) now requires `fsdp_version: 2` with `state_dict_type: FULL_STATE_DICT`; EP with tensor parallelism is rejected.

WAS THIS USEFUL? 👍 👎

Intel AutoRound

Feedback ⓘ v0.16.0 NOTES ↗ RANK

AutoRound quantizes language and multimodal models through weight-rounding optimization and exports them for inference engines.

> AutoRound 0.16.0 introduces AutoScheme for automated mixed-bit quantization recipe generation, adds new low-bit formats (MXFP4, NVFP4, FP8) and GGUF export, extends CUDA weight-only quantization to 5/6/7-bit via a new Humming Kernel, and broadens deployment to Gaudi, Intel XPU, vLLM and SGLang, alongside calibration, CLI and model-free compression workflow improvements.

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

01 **reference_output parameter in quantize_block** NEW 80

Adds a `reference_output` parameter to `quantize_block` in the orchestrator to skip the `collect_reference` step.

— Names exact parameter, function and skipped step. v0.16.0

02 **New quantization data type support** NEW 75

Adds support for `MXFP4`, `NVFP4`, and `FP8` data types alongside existing `WNA16` quantization schemes; extends this with experimental `NVFP4 E5M3` support for model-free quantization, and `FP8` model loading with a ragged-tile dequantization fallback.

— Names formats and a fallback mechanism, lacks usage example. v0.16.0

03 **AutoScheme mixed-bit quantization recipes** NEW 70

`AutoScheme` generates mixed-bit quantization recipes in minutes, with usage documented in `docs/step_by_step.md#autoscheme`; per-layer sensitivity scoring is now streamed block-by-block to reduce peak memory during recipe generation.

— Names doc location and memory mechanism, no runnable example. v0.16.0

04 **5/6/7-bit WOQ CUDA quantization via Humming Kernel** NEW 70

Supports 5, 6, and 7 weight-only quantization (WOQ) bits on CUDA in both Transformers and vLLM backends via the Humming Kernel.

— Names kernel and backends, but no command shown. v0.16.0

THINNER COVERAGE BELOW

05 **vLLM, SGLang and GGUF integration** NEW 55

Integrates with `vLLM` and `SGLang` inference ecosystems alongside existing Transformers support, and adds `GGUF` low-bit model format for export/deployment.

— Lists integrations and format, no usage detail. product docs

06 **Model-free compression workflow improvements** IMPROVED 55

Adds resume functionality for model-free compression, allowing interrupted quantization jobs to be continued; adds `WOQ` fake quantization support in model-free mode; and adds transformers checkpoint conversion mapping support for missing tensors.

— Describes three workflow additions without commands. v0.16.0

07 **Calibration dataset flexibility** IMPROVED 45

Adds `fineweb-edu` dataset as a calibration backup option, and allows non-MLLM calibration datasets when not quantizing vision components.

— Names dataset option, lacks concrete usage. v0.16.0

08 **Gaudi and Intel XPU deployment support** NEW 40

Supports deployment on Gaudi accelerators and Intel XPUs alongside existing Intel CPU and NVIDIA GPU targets.

— Names hardware targets only, no setup detail. product docs

09 **CLI export and reproducibility options** NEW 40

Supports custom max shard size in CLI and model export, and adds deterministic algorithms flags support in the CLI.

— Describes CLI additions without exact flag names. v0.16.0

10 **MoE model tuning speedup** IMPROVED 35

Delivers a 2.5–6× speedup for MoE model tuning.

— Gives concrete speedup figure, no mechanism explained. v0.16.0

11 **Opt-in diffusion tuning cache prefetch** NEW 25

Adds opt-in diffusion tuning cache prefetch.

— Bare mention, no mechanism or flag given. v0.16.0

v0.16.0 NOTES ↗ LATEST STABLE

AutoRound v0.16.0 adds 5/6/7-bit WOQ CUDA support, resume for model-free compression, NVFP4, FP8 loading, and 2.5–6× MoE tuning speedup.

└──▷ GET THIS VERSION

```
$ git clone --branch v0.16.0 https://github.com/intel/auto-round.git
# already have the repo? check out this version:
$ git checkout v0.16.0
```

Supports 5, 6, and 7 weight-only quantization (WOQ) bits on CUDA in Transformers and vLLM backends via the Humming Kernel.

Implements resume functionality for model-free compression, allowing interrupted quantization jobs to be continued.

Adds experimental NVFP4 E5M3 support and related components for model-free quantization.

Adds FP8 model loading and ragged-tile dequantization fallback.

Adds `fineweb-edu` dataset as a calibration backup option.

+9 more

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

Comfy Org ComfyUI

Feedback ⓘ v0.38.0 NOTES ↗ RANK

ComfyUI is a node-graph interface, API, and backend for running diffusion models, chaining generation steps into reusable visual workflows.

> ComfyUI's v0.38.0 release and docs updates add a built-in Comfy MCP assistant for building and running workflows (now GA on Cloud), Seedance 2.5 Draft mode for fast 480p-to-1080p previews, and broad new model/partner-node support (Ming-Image, Hunyuan Image 3.5, GPT-6 Sol/Luna, Seedream 5.0 Flash, ID-V2V Wan 2.1, w6a8 quantization), while removing the Sora video nodes and dropping the torchaudio dependency.

└──▷ WHAT SHIPPED · 20 FEATURES most completely described first ? what's the number?

01 Seedance 2.5 Draft mode for previews NEW 80

Seedance 2.5 gains a Draft mode that produces a fast 480p preview via a new `draft_task_id` field, which feeds directly into the full 1080p render job; delivered via ByteDance partner nodes.

Use Seedance 2.5 Draft mode to get a fast 480p preview and then pipe its draft_task_id into the full 1080p render job.

📍 In the ComfyUI canvas, add the Seedance 2.5 node, enable Draft mode, run the graph to get the 480p preview, then connect the `draft_task_id` output to the final 1080p Seedance render node and queue again.

— Names field and preview-to-final workflow, with a usage example. v0.38.0

02 JsonExtractString prefix/suffix handling IMPROVED 75

The `JsonExtractString` node now handles JSON payloads wrapped in leading or trailing prefix/suffix text, such as a model response like `Result: {"key":"value"}`, by setting a prefix/suffix string to strip before parsing.

Parse a JSON payload that arrives wrapped in a prefix or suffix string, such as a model response like `Result: {"key":"value"}`.

📍 In the ComfyUI canvas, add the `JsonExtractString` node, connect your raw text output to its input, set the prefix to 'Result: ' and leave the suffix empty, then wire the parsed output to the next node.

— Named node with worked example of exact parsing behavior. v0.38.0

03 HDR color space conversion NEW 60

Adds `HDR LogC3` and `ACEScct` color space options to the Convert Color Space node, letting a workflow convert footage into these targets for HDR-aware compositing.

Convert footage to the ACEScct color space for HDR-aware compositing inside a ComfyUI workflow.

> 📍 Add a `Convert Color Space` node to your workflow, set the source space to match your input (e.g. sRGB), and select `ACEScct` or `HDR LogC3` from the target color space dropdown.

— Named node and formats with a usage example. v0.38.0

---

04 **TextGenerate system_prompt and thinking output** NEW

60

Adds a `system_prompt` input and a separate `thinking` output to the `TextGenerate` node, letting a workflow steer model behaviour without polluting the user prompt and route the model's reasoning to a separate socket.

Use the new system_prompt slot in a TextGenerate workflow to steer model behaviour without polluting the user prompt.

> 📍 In your workflow, open the TextGenerate node, connect a text string to the new `system_prompt` input, and wire the separate `thinking` output socket to a preview or downstream node.

— Named fields with example wiring steps. v0.38.0

THINNER COVERAGE BELOW

05 **Ming-Image model support** NEW

50

Adds Ming Image 0.1, a 6B parameter text-to-image model aimed at text-heavy design layouts with transparent RGBA output, plus model loading and inference support for Ming-Image.

— Names model and specs but no usage steps. v0.38.0

---

06 **Per-block attention selection for Lumina** NEW

50

Adds `comfy_attention` per-block attention selection for Lumina family models, letting model files specify which attention implementation is used per block, for lower memory usage.

— Names the config surface but no usage example. v0.38.0

---

07 **New partner model integrations** NEW

45

Adds Hunyuan Image 3.5 text-to-image and image-edit partner nodes (Tencent), GPT-6 Sol and Luna models to the OpenAI partner node, and Seedream 5.0 Flash via ByteDance

partner nodes.

— Lists named models but no usage detail. v0.38.0

08 **Additional model and quantization support** NEW 45

Adds support for ID-V2V Wan 2.1/VACE-based video models, MiniMax-H3 Fun-Controlnet-Union 2.0, Qwen-Image 2.1 union fun ControlNet, and the `w6a8` quantization format.

— Names four surfaces but no usage detail. v0.38.0

09 **Torchaudio dependency dropped** BREAKING 40

Removes `torchaudio` as a required dependency; workflows or custom nodes relying on it will need to supply it independently.

— Names the removed dependency and notes migration need. v0.38.0

10 **Comfy MCP workflow assistant** NEW 40

Adds Comfy MCP, a built-in ComfyUI assistant for building, editing, and running workflows, now generally available on Cloud; local Comfy Desktop connection is coming soon.

— Describes assistant with setup screenshots but no config detail. product docs

11 **Qwen-Image 2.1 tiny VAE** NEW 35

Adds a tiny VAE for Qwen-Image 2.1 for faster previews and lower memory use.

— Thin one-line addition, no mechanism given. v0.38.0

12 **Sora video nodes removed** BREAKING 30

The Sora video nodes have been removed from ComfyUI; any workflow using them will break.

— Names removed nodes but gives no migration path. product docs

13 **CUDA graphs on ACE Step 1.5** IMPROVED 30

Enables CUDA graphs and memory compiler on the ACE Step 1.5 autoregressive model.

— Names mechanism but no usage steps. v0.38.0

14 **H3 VAE tile blending** IMPROVED 30

Blends H3 VAE tiles against composited neighbours for smoother tiled VAE decoding.

— Names mechanism briefly but no usage detail. v0.38.0

15 **Non-blocking asset scans** IMPROVED 25

Asset scans and output rescans no longer block the UI on large libraries; output folders are now listed incrementally during asset scans to keep the UI responsive.

— Describes the behavior change but no concrete usage. v0.38.0

16 **RDNA2 GPU architecture support** NEW 25

Adds missing RDNA2 GPU architectures to the supported hardware list.

— Bare hardware list addition, no further detail. v0.38.0

17 **Yue2 autoregressive speedup** IMPROVED 20

Adds Yue2 AR for faster autoregressive generation on Yue2, and further speeds up Yue2 autoregressive inference.

— Bare performance claim, no mechanism or numbers. v0.38.0

18 **NPU async weight offload** NEW 20

Adds async weight offload stream support for NPU devices.

— Single line, no mechanism or usage detail. v0.38.0

19 **Fast disk detection for model loaders** IMPROVED 20

Adds fast disk detection to all model loaders.

— Bare one-line addition, no mechanism given. v0.38.0

20 **Qwen 3.5/3.8 long-context speedup** IMPROVED 20

Speeds up generation on Qwen 3.5/3.8 with long contexts.

— Bare performance claim, no numbers or mechanism. v0.38.0

└──▷ **BREAKING ON UPGRADE**

! The Sora video nodes have been removed from ComfyUI; any workflow using them will break.

! The torchaudio dependency has been dropped; workflows or custom nodes relying on it will need to supply it independently.

v0.38.0 NOTES ↗ LATEST STABLE

ComfyUI v0.38.0 adds Hunyuan Image 3.5, Ming-Image, ID-V2V Wan 2.1, w6a8 quantization, HDR color spaces, and NPU async weight offload.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.38.0 https://github.com/Comfy-Org/ComfyUI.git
# already have the repo? check out this version:
$ git checkout v0.38.0
```

└──▷ **HOW TO FIND IT**

Use the new system_prompt slot in a TextGenerate workflow to steer model behaviour without polluting the user prompt.

> 📍 In your workflow, open the TextGenerate node, connect a text string to the new `system_prompt` input, and wire the separate `thinking` output socket to a preview or downstream node.

Convert footage to the ACEScct color space for HDR-aware compositing inside a ComfyUI workflow.

> 📍 Add a `Convert Color Space` node to your workflow, set the source space to match your input (e.g. sRGB), and select `ACEScct` or `HDR LogC3` from the target color space dropdown.

Start ComfyUI without failing when asset packages are absent — useful in minimal CI or container environments.

```
$ comfyui --enable-assets
```

Adds `HDR LogC3` and `ACEScct` color space options to the Convert Color Space node.

Supports `.comfy_attention` per-block attention selection for Lumina family models.

Adds `system_prompt` input and a separate thinking output to the `TextGenerate` node.

Adds Hunyuan Image 3.5 text-to-image and image-edit partner nodes (Tencent).

Adds GPT-6 Sol and Luna models to the OpenAI partner node.

+19 more

WAS THIS USEFUL? 👍 👎

mlxtop

Feedback ⓘ v1.2.0 NOTES ↗ RANK

The mlxtop terminal UI monitors local LLMs on Apple Silicon, showing loaded models, memory use, GPU activity, and generation speed.

> mlxtop expands beyond Apple Silicon with native Linux support and a full multi-GPU NVIDIA dashboard, and adds interactive chart controls, a slowdown-evidence diagnostics view, richer prompt-load and oMLX distributed-request tracking, and a pinnable installer.

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

01 Multi-GPU NVIDIA dashboard in GPU DEVICES panel NEW

95

On Linux, mlxtop detects NVIDIA GPUs via `nvidia-smi` and shows a per-card overview with driver index, name, utilization, VRAM used/total, and temperature in the GPU DEVICES panel; wide terminals add a VRAM bar and textual load state column. Cards are tracked by UUID so selections follow the same physical device if driver indices change, a GPU max chart shows the highest utilization across all cards rather than an average, and cards are navigated with `[` / `]`. The `--once` flag lists all detected NVIDIA cards and their UUIDs without pagination.

Enumerate all NVIDIA GPUs and their stable UUIDs in one shot — useful for scripting or confirming device identity before a monitoring session.

```
$ mlxtop --once
```

Dump all detected NVIDIA cards with their UUIDs for scripting or auditing GPU inventory without entering the live dashboard.

```
$ mlxtop --once
```

— Names panel, flag, keybinds and detection mechanism fully v1.2.0

02 MLXTOP_VERSION installer pinning NEW

80

The terminal installer supports the `MLXTOP_VERSION` environment variable to pin installation to a specific release instead of always installing the latest, e.g. `MLXTOP_VERSION=1.2.0 curl -fsSL https://raw.githubusercontent.com/maximpri/mlxtop/main/scripts/install.sh | sh`.

Pin the installer to this exact release so CI or shared machines always get v1.2.0 rather than the latest.

```
$ MLXTOP_VERSION=1.2.0 curl -fsSL
https://raw.githubusercontent.com/maximpri/mlxtop/main/scripts/in
stall.sh | sh
```

Pin the installer to a specific mlxtop release on a Linux server rather than always pulling the latest build.

```
$ MLXTOP_VERSION=1.1.2 curl -fsSL
https://raw.githubusercontent.com/maximpri/mlxtop/main/scripts/in
stall.sh | sh
```

— Exact env var and runnable install command provided `v1.1.2` `v1.2.0`

03 **Interactive chart and dashboard controls** NEW 75

Charts can be selected with Tab, arrow keys, or click; zoomed in history with `+` / `-` or the mouse wheel; enlarged or restored with Enter / Esc. The sampling/refresh interval is changed with `{` / `}`, adjustable between 1 and 60 seconds, and Ctrl-Tab switches between dashboard views.

— Every keybind named but no runnable example given `v1.2.0`

04 **Prompt load panel enhancements** IMPROVED 70

The prompt load display now shows exact input-token counts (including cached tokens), request freshness, and the recent median and range; the bottom border of the prompt load chart shows median and range statistics from up to eight contiguous observations. On wide terminals (160 columns) the prompt load panel moves beside the generation and prefill charts in the first chart row instead of sitting in a separate row.

— Detailed layout and stats but no interactive control named `v1.2.0`

05 **Static Linux binaries for x86_64 and aarch64** NEW 60

mlxtop ships static Linux binaries for x86_64 and aarch64, built as musl binaries with no runtime library dependencies, extending platform support beyond Apple Silicon Macs.

— Platforms and build type named, no install command tied here `v1.2.0` `v1.1.2`

THINNER COVERAGE BELOW

06 **oMLX distributed request tracking** IMPROVED 55

oMLX queue, rate, and output readings now cover every loaded model and every concurrent request rather than a single active one. Distributed requests are tracked

separately in prompt history using rank zero's request IDs instead of the shared `rank0` placeholder, and the oMLX prompt history is upgraded to a full panel with one bar per request and UTC timestamps.

— Mechanism named but no UI path or command given v1.1.2

07 **Terminal bell for critical memory pressure** NEW 50

mlxtop rings a terminal bell alert when it detects critical memory pressure or severe paging; the alert is acknowledged with the `a` key.

— Clear trigger and keybind, but no example shown v1.2.0

08 **Diagnostics view for measured slowdowns** NEW 40

A new diagnostics view shows measured slowdowns with supporting evidence; high GPU utilization alone no longer triggers an alarm.

— Behaviour change described but no view path or metric named v1.2.0

09 **LM Studio Bionic runtime detection** NEW 25

mlxtop detects LM Studio's Bionic app as a supported runtime.

— Bare one-line addition with no further detail v1.1.2

v1.2.0 NOTES ↗ LATEST STABLE

mlxtop 1.2.0 adds multi-GPU support, interactive chart controls, and smarter diagnostics for Linux LLM workloads.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v1.2.0 https://github.com/maximpri/mlxtop.git
# already have the repo? check out this version:
$ git checkout v1.2.0
```

└──▷ **TRY IT**

Pin the installer to this exact release so CI or shared machines always get v1.2.0 rather than the latest.

```
$ MLXTOP_VERSION=1.2.0 curl -fsSL
https://raw.githubusercontent.com/maximpri/mlxtop/main/scripts/insta
ll.sh | sh
```

Adds per-card NVIDIA GPU overview with utilization, VRAM, and temperature; navigate cards with `[` / `]` keys.

Adds interactive chart controls: select charts with Tab, arrow keys, or click; zoom history with `+` / `-` or mouse wheel; enlarge/restore with Enter / Esc; change sampling interval with `{` / `}`.

Adds terminal bell alert for critical memory pressure and severe paging, acknowledged with `a` key.

Adds exact input-token counts (including cached tokens), request freshness, recent median, and range to prompt load display.

Adds diagnostics view showing measured slowdowns with supporting evidence; high GPU utilization alone no longer triggers an alarm.

+1 more

v1.1.2 NOTES ↗ STABLE

mlxtop v1.1.2 adds Linux x86_64/aarch64 binaries, a pinnable installer via `MLXTOP_VERSION`, and improved oMLX distributed-request tracking.

└──▷ **GET THIS VERSION**

```
$ git clone --branch v1.1.2 https://github.com/maximpri/mlxtop.git
# already have the repo? check out this version:
$ git checkout v1.1.2
```

└──▷ **TRY IT**

Pin the installer to a specific mlxtop release on a Linux server rather than always pulling the latest build.

```
$ MLXTOP_VERSION=1.1.2 curl -fsSL
https://raw.githubusercontent.com/maximpri/mlxtop/main/scripts/insta
ll.sh | sh
```

Adds `MLXTOP_VERSION` environment variable to the terminal installer to pin a specific release instead of always installing the latest.

Ships static Linux binaries for x86_64 and aarch64, extending platform support beyond Apple Silicon Macs.

Expands oMLX queue, rate, and output readings to cover every loaded model and every concurrent request, not just a single active one.

Tracks oMLX distributed requests separately in prompt history using rank zero's request IDs instead of the shared `rank0` placeholder.

Detects LM Studio's Bionic app as a supported runtime.

+1 more

WAS THIS USEFUL? 👍 👎

vMLX Feedback ⓘ v1.6.72 NOTES ↗ RANK

The vMLX server runs compressed MLX models on Apple Silicon with disk caching, paged memory, continuous batching, and hybrid SSM scheduling.

Exposes backend cache/prefill metrics via Ollama; preserves tool arguments, reasoning history, and logprob metadata; cleans up canceled streams; adds UI prefill-counter i18n.

└──▷ GET THIS VERSION

```
$ git clone --branch v1.6.72 https://github.com/jjang-ai/vmlx.git
# already have the repo? check out this version:
$ git checkout v1.6.72
```

Exposes backend cache and prefill metrics through the Ollama API endpoint.

Preserves tool arguments and reasoning history across API and chat turns.

Retains logprob metadata for hidden parser markers in API and chat responses.

Cleans up canceled streams in the API and chat layer.

Translates the measured-prefill counter consistently across supported UI languages.

NEW SUPPORTED PLATFORM Naive N0.5 runtime and caching support

NEW SUPPORTED PLATFORM GLM 5.3 Flash runtime and caching support

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

Pinecone

Feedback ⓘ snapshot-20260930 NOTES ↗ RANK

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

> Pinecone introduced namespace aliases in its serverless indexes, giving a full CRUD API to create movable pointers that let applications repoint reads to new data atomically with zero downtime.

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

01 Namespace aliases for zero-downtime repointing NEW

95

Pinecone adds a namespace alias API for serverless indexes: `GET /indexes/{index}/namespaces/aliases` lists aliases with filtering by name prefix or target namespace, a `limit` parameter, and `paginationToken` / `pagination.next` pagination (up to 100 per page); `POST /indexes/{index}/namespaces/aliases` creates an alias resolving reads to an existing target namespace (writes to an alias name are rejected); `GET /indexes/{index}/namespaces/aliases/{alias}` describes an alias and its target; `PATCH /indexes/{index}/namespaces/aliases/{alias}` atomically repoints an alias to a new target namespace so reads arriving after the response are served from the new namespace with no polling required; and `DELETE /indexes/{index}/namespaces/aliases/{alias}` removes an alias without affecting the target namespace or its data.

Atomically swap a live namespace alias to a freshly ingested namespace — zero-downtime data refresh with no code change.

```
$ curl -X PATCH 'https://your-index-host/indexes/my-index/namespaces/aliases/prod-alias' \
  -H 'Api-Key: <PINECONE_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"target_namespace": "vectors-2026-10-01"}'
```

Page through all namespace aliases to audit which namespaces are currently live behind each alias name.

```
$ curl 'https://your-index-host/indexes/my-
index/namespaces/aliases?limit=100' \
  -H 'Api-Key: <PINECONE_API_KEY>'
```

Create a new namespace alias pointing at an existing namespace so applications can query it by stable name without knowing the underlying namespace.

```
$ curl -X POST 'https://your-index-host/indexes/my-
index/namespaces/aliases' \
  -H 'Api-Key: <PINECONE_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"name": "prod-alias", "target_namespace": "vectors-2026-
09-30"}'
```

— Full endpoint set with methods, params and behaviour specified. snapshot-20260930

snapshot-20260930 NOTES ↗ LATEST STABLE

Pinecone adds namespace aliases — movable pointers that let you atomically repoint reads to new data with no downtime or redeploy.

└──▷ TRY IT

Atomically swap a live namespace alias to a freshly ingested namespace — zero-downtime data refresh with no code change.

```
$ curl -X PATCH 'https://your-index-host/indexes/my-
index/namespaces/aliases/prod-alias' \
  -H 'Api-Key: <PINECONE_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"target_namespace": "vectors-2026-10-01"}'
```

Page through all namespace aliases to audit which namespaces are currently live behind each alias name.

```
$ curl 'https://your-index-host/indexes/my-index/namespaces/aliases?
limit=100' \
  -H 'Api-Key: <PINECONE_API_KEY>'
```

Create a new namespace alias pointing at an existing namespace so applications can query it by stable name without knowing the underlying namespace.

```
$ curl -X POST 'https://your-index-host/indexes/my-index/namespaces/aliases' \
  -H 'Api-Key: <PINECONE_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"name": "prod-alias", "target_namespace": "vectors-2026-09-30"}'
```

Adds `GET /indexes/{index}/namespaces/aliases` endpoint to list namespace aliases in a serverless index, with filtering by name prefix or target namespace, `limit` parameter, and `paginationToken` / `pagination.next` for pagination (up to 100 aliases per page).

Adds `POST /indexes/{index}/namespaces/aliases` endpoint to create a namespace alias that resolves reads to an existing target namespace; writes to an alias name are rejected.

Adds `GET /indexes/{index}/namespaces/aliases/{alias}` endpoint to describe a namespace alias and the namespace it resolves to.

Adds `PATCH /indexes/{index}/namespaces/aliases/{alias}` endpoint to atomically repoint an alias to a new target namespace; reads arriving after the response are served from the new namespace with no polling required.

Adds `DELETE /indexes/{index}/namespaces/aliases/{alias}` endpoint to remove an alias without affecting the target namespace or its data.

WAS THIS USEFUL? 👍 👎

Qdrant

Feedback ⓘ announced · Sep 29, 2026 BLOG ↗ RANK

Qdrant - High-performance, massive-scale Vector Database and Vector Search Engine for the next generation of AI. Also available in the cloud https://cloud.qdrant.io/

> Qdrant released a research preview called Constella, introducing hot-swappable query encoder models that let developers change search speed/accuracy tradeoffs without re-embedding existing collections.

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

01 **Constella swappable query encoders** NEW 88

Qdrant added `Qdrant/constella-zero` and `Qdrant/constella-nano` query encoder models on Hugging Face, which can be hot-swapped against existing `Qdrant/stella-en-400M-v5-doc-onnx` document vectors in the same Qdrant collection without re-embedding documents. `Qdrant/constella-zero` uses token lookup + pooling (0.081 ms warm p50) while `Qdrant/constella-nano` is a 34.5M-parameter transformer (3.13 ms warm p50); either can be passed per-request into TextEmbedding(...) while document vectors stay fixed. `Qdrant/constella-zero` can also be paired with BM25 for hybrid search without running a transformer per query. The preview installs via `pip install 'fastembed @ git+https://github.com/Dylancouzon/fastembed.git@constella-research-preview'` on the `constella-research-preview` branch of FastEmbed.

Encode documents once with Stella, then query the same collection with the ultra-fast Zero model — swap to Nano by changing one model name.

```python
from fastembed import TextEmbedding
from qdrant_client import QdrantClient, models

client = QdrantClient(":memory:")
client.create_collection(
    "documents",
    vectors_config=models.VectorParams(size=1024,
distance=models.Distance.COSINE),
)

documents = [
    "Solar panels convert sunlight into electricity.",
    "Wind turbines generate electricity from moving air.",
]
stella = TextEmbedding("Qdrant/stella-en-400M-v5-doc-onnx")
vectors = stella.embed(documents)
client.upsert(
    "documents",
    points=[
        models.PointStruct(id=i, vector=vector.tolist(), payload=
{"text": text})
        for i, (text, vector) in enumerate(zip(documents,
vectors))
    ],
)

# Switch to Nano: change "constella-zero" -> "constella-nano"
query_model = TextEmbedding("Qdrant/constella-zero")
query_vector = next(query_model.embed(["How can we get energy
from the sun?"]))
results = client.query_points("documents",
query=query_vector.tolist(), limit=2).points
for result in results:
    print(result.payload["text"])
```

Install the Constella research-preview branch of FastEmbed alongside qdrant-client to access the new models.

```
$ pip install "fastembed @
git+https://github.com/Dylancouzon/fastembed.git@constella-
research-preview" qdrant-client
```

— Names exact models, latencies, install command and code path `launch-20260929-e233f6f4`

announced · Sep 29, 2026 BLOG ↗

Constella research preview: swap query encoders (Zero, Nano, full Stella) without re-embedding your Qdrant collection

└──▷ **USE IT**

Encode documents once with Stella, then query the same collection with the ultra-fast Zero model — swap to Nano by changing one model name.

python

```python
from fastembed import TextEmbedding
from qdrant_client import QdrantClient, models

client = QdrantClient(":memory:")
client.create_collection(
    "documents",
    vectors_config=models.VectorParams(size=1024,
distance=models.Distance.COSINE),
)

documents = [
    "Solar panels convert sunlight into electricity.",
    "Wind turbines generate electricity from moving air.",
]
stella = TextEmbedding("Qdrant/stella-en-400M-v5-doc-onnx")
vectors = stella.embed(documents)
client.upsert(
    "documents",
    points=[
        models.PointStruct(id=i, vector=vector.tolist(), payload=
{"text": text})
        for i, (text, vector) in enumerate(zip(documents, vectors))
    ],
)

# Switch to Nano: change "constella-zero" -> "constella-nano"
query_model = TextEmbedding("Qdrant/constella-zero")
query_vector = next(query_model.embed(["How can we get energy from
the sun?"]))
results = client.query_points("documents",
query=query_vector.tolist(), limit=2).points
for result in results:
    print(result.payload["text"])
```

Install the Constella research-preview branch of FastEmbed alongside qdrant-client to access the new models.

```
$ pip install "fastembed @
git+https://github.com/Dylancouzon/fastembed.git@constella-research-
preview" qdrant-client
```

Adds `Qdrant/constella-zero` and `Qdrant/constella-nano` query encoder models on Hugging Face, hot-swappable against `Qdrant/stella-en-400M-v5-doc-onnx` document vectors in the same Qdrant collection without re-embedding

Installs via `pip install 'fastembed @ git+https://github.com/Dylancouzon/fastembed.git@constella-research-preview'` on the `constella-research-preview` branch of FastEmbed

Enables per-request query model selection — pass `Qdrant/constella-zero` (token lookup + pooling, 0.081 ms warm p50) or `Qdrant/constella-nano` (34.5M-parameter transformer, 3.13 ms warm p50) into TextEmbedding(...) while document vectors remain fixed

Supports hybrid search by pairing `Qdrant/constella-zero` with BM25 in Qdrant without running a transformer for each query

---

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

UK AI Security Institute Inspect AI

Feedback ⓘ

0.3.273 CODE ↗ RANK

Inspect is the UK AI Security Institute's framework for large language model evaluations, covering datasets, solvers, scorers, and agentic eval tasks.

Adds stable-prefix prompt caching for Anthropic and explicit cache boundaries for OpenAI GPT-5.6+; faster `resolve_attachments` reads for long conversations.

└──▷ GET THIS VERSION

```
$ git clone --branch 0.3.273
https://github.com/UKGovernmentBEIS/inspect_ai.git
# already have the repo? check out this version:
$ git checkout 0.3.273
```

Enables Anthropic prompts to mark a stable prefix for caching while allowing the remaining text to change freely on each call.

Enables explicit prompt-cache boundaries for OpenAI GPT-5.6 and later; older models retain automatic caching.

Speeds up reading long conversations when using `resolve_attachments`; in full mode, deeply nested model API call content may retain two additional nesting levels.

NEW SUPPORTED PLATFORM GPT-6.1 Sol model support

UPDATED SUPPORTED PLATFORM Mistral provider updated to mistralai 3.0 (httpx2)

└──▷ BREAKING ON UPGRADE

! mistralai 3.0 is now the minimum required version for the Mistral provider; installations using an older mistralai package will break.

WAS THIS USEFUL? 👍 👎

Arize Phoenix

Feedback ⓘ @arizeai/phoenix-cli@1.18.6 NOTES ↗

RANK

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

> Arize Phoenix added Qdrant vector database tracing, prompt version tag management helpers in the client, and keyboard navigation improvements to the PXI slash-command menu.

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

01 **Prompt version tag helpers in client** NEW 80

The `@arizeai/phoenix-client` adds `upsertPromptVersionTag` and `deletePromptVersionTag` helpers to the `prompts` subpath for creating, moving, and deleting prompt-scoped version tags.

— Names exact helper functions and subpath @arizeai/phoenix-client@7.15.0

02 **PXI slash-command menu navigation** IMPROVED 75

The PXI slash-command hint menu gains arrow-key navigation (wrapping at both ends), Tab completion, and Enter-to-run, so typing `/he` and pressing Enter executes `/help`; a bare `/` lists all commands with `/help` first.

— Describes exact interaction behavior and example commands @arizeai/phoenix-cli@1.18.6

THINNER COVERAGE BELOW

03 **Qdrant vector database integration** NEW 45

Adds Qdrant vector database integration, enabling trace retrieval from Qdrant to identify which stage of a search pipeline needs fixing.

— Names integration but no config or setup detail product docs

@arizeai/phoenix-cli@1.18.6 NOTES ↗ LATEST STABLE

PXI slash-command hints now support arrow-key navigation, Tab completion, and Enter to run.

└──▷ **GET THIS VERSION**

```
$ git clone --branch @arizeai/phoenix-cli@1.18.6
https://github.com/Arize-ai/phoenix.git
# already have the repo? check out this version:
$ git checkout @arizeai/phoenix-cli@1.18.6
```

`@arizeai/phoenix-client@7.15.0` NOTES ↗ STABLE

Arize Phoenix client adds `upsertPromptVersionTag` and `deletePromptVersionTag` helpers for managing prompt version tags.

└──▷ **GET THIS VERSION**

```
$ git clone --branch @arizeai/phoenix-client@7.15.0
https://github.com/Arize-ai/phoenix.git
# already have the repo? check out this version:
$ git checkout @arizeai/phoenix-client@7.15.0
```

Adds `upsertPromptVersionTag` and `deletePromptVersionTag` helpers to the `prompts` subpath for creating, moving, and deleting prompt-scoped version tags.

WAS THIS USEFUL? 👍 👎

Braintrust

Feedback ⓘ snapshot-20260930 NOTES ↗ RANK

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

> Braintrust's biggest additions this window are external MCP server connections in Loop (with OAuth 2.0 PKCE securing its own MCP server) and a new `bt custom-views` CLI for building and publishing trace/dataset views, alongside Patterns automation improvements, AI provider model governance, and new tracing integrations for Claude Code, Grok, and Google Antigravity.

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

01 bt custom-views CLI for trace and dataset views NEW

100

A `bt custom-views` CLI adds `bt custom-views bootstrap '<View Name>'` and `bt custom-views dataset bootstrap '<View Name>'` to scaffold starter `.tsx` files in `braintrust-custom-views/`; `bt custom-views preview <file> --url <BRAINTRUST_TRACE_URL>` to hot-reload a trace view against a live trace; `bt custom-views dataset preview <file> --project <name> --dataset <name> --row-index <n>` (or `--row-id <ROW_ID>`) to preview against a specific row; and `bt custom-views push <directory> --project <name>` to publish, with `--if-exists replace` to overwrite a view matched by `slug`. Views are defined with customTraceView() and customDatasetView() helpers from `braintrust/custom-views`, and datasets can be identified by name (`dataset: { name: 'my-dataset' }`) or ID (`dataset: { id: '<DATASET_ID>' }`). Requires SDK v3.33.0+ and ships as `@braintrust/bt` on npm, independent of the JS SDK, installable via `npm install --save-dev @braintrust/bt` or `npm install -g @braintrust/bt`.

Scaffold and push a trace view from local files so your team can version-control and iterate on it before publishing.

```
$ bt custom-views bootstrap 'Trace Review'
# edit braintrust-custom-views/trace-review.trace-view.tsx
bt custom-views push ./braintrust-custom-views --project my-project
```

Hot-reload a trace view locally against a real trace URL before pushing, to verify layout and span selection logic.

```
$ bt custom-views preview ./braintrust-custom-views/trace-review.trace-view.tsx \
  --url "<BRAINTRUST_TRACE_URL>" \
  --project my-project
```

Preview a dataset view against a specific row, then push and replace the existing published view.

```
$ bt custom-views dataset preview ./braintrust-custom-views/dataset-review.dataset-view.tsx \
  --project my-project \
  --dataset my-dataset \
  --row-index 0
# when satisfied:
bt custom-views push ./braintrust-custom-views --project my-project --if-exists replace
```

Bootstrap, preview, and publish a custom view from the CLI without touching the JavaScript SDK.

```
$ npm install -g @braintrust/bt
bt custom-views bootstrap my-view
bt custom-views preview my-view
bt custom-views push my-view
```

— Full runnable commands, flags, helpers and install instructions. product docs

02 Project search and metrics performance settings NEW

80

Adds four project settings: Enable log search optimization builds a full-text index to speed up text-based log filter queries; Enable shingled search optimization indexes multi-word phrases to speed up phrase filter queries; Enable metric columnstore writes trace-level metrics in columnar format to speed up trace queries; `Use metrics.start for monitor charts time bucketing` controls how monitor chart time buckets are anchored.

Enable full-text and phrase search indexes on a high-volume project so that text-based log filters return results faster.

📍 In the project Settings page, turn on Enable log search optimization, then turn on Enable shingled search optimization to also cover multi-word phrase queries.

Speed up the trace metric queries that power dashboards by switching to columnar storage for trace-level metrics.

> 📍 In the project Settings page, turn on `Enable metric columnstore`.

— Names all four toggles and where to enable each. `product docs`

---

03 **External MCP connections in Loop** NEW 75

External MCP connections in Loop launched as a public preview on Braintrust-hosted deployments: shared OAuth connections or API keys can be configured per project in Loop's project MCP settings, letting Loop use an external server's tools and data alongside Braintrust traces. Project MCP credentials are encrypted and stored on the data plane using the same mechanism as function environment variables, decrypted by Loop's runtime at request time; scheduled Loop runs, including Patterns automations, can invoke all tools allowed by the shared credentials without per-run interactive approval. Requires data plane v2.16.0+.

Set up a shared external MCP connection for a project so all members and scheduled Patterns automations can use it without per-user approval.

> 📍 In Loop, go to your project's MCP settings, then add a shared OAuth connection or API key under `External MCP connections`. Scope the credential to only the access you intend to share with the project.

— Names encryption mechanism, version requirement, and settings location. `snapshot-20260930`

---

04 **AI provider model governance and browsing** NEW 60

Adds model restrictions under AI provider governance to control which models each team can access, including a `Selected models` option for project providers on Braintrust-hosted deployments. Adds a `Models` tab in AI provider setup to browse available models and copy a Gateway command to test one.

— Names the exact settings and tab but no CLI command. `snapshot-20260930`

THINNER COVERAGE BELOW

---

05 **bt CLI tracing for Grok and Google Antigravity** NEW 55

Adds `bt` CLI (v0.19.3+) support for tracing Grok and Google Antigravity coding sessions as Braintrust traces, capturing prompts, model calls, and tool activity without interrupting the agent if tracing fails.

— Names CLI version and providers, no exact command shown. `snapshot-20260930`

---

06 **OAuth 2.0 PKCE for Braintrust MCP server** NEW 50

Adds OAuth 2.0 with PKCE (Proof Key for Code Exchange) authentication for clients connecting to the Braintrust MCP server, issuing access tokens with refresh capabilities for secure, long-lived sessions.

— Explains mechanism but gives no setup steps. product docs

07 **Claude Code OpenTelemetry trace mapping** NEW 50

Maps Claude Code's native OpenTelemetry traces and events to LLM and tool spans with token metrics, cost, and conversation messages (requires data plane v2.16.0+).

— Describes mapping and requirement but no usage steps. snapshot-20260930

08 **Loop thread sharing** NEW 45

Adds Loop thread sharing so project members can review or continue an investigation together with view-only or chat access (requires data plane v2.15.0+).

— States capability and requirement but no navigation path. snapshot-20260930

└──▷ **ALSO FROM THESE RELEASES**

Deduplicate SQL query results to find the unique combinations of model and status across your traces.

```
$ SELECT DISTINCT model, status FROM traces WHERE project_id =
'<project_id>'
```

└──▷ **BREAKING ON UPGRADE**

! On Braintrust-hosted deployments, project providers now override organization providers only for models they serve; organization providers can serve other models using their own credentials. If you rely on project providers to restrict model access or spending for all models, review organization provider model selections and permissions.

snapshot-20260930 NOTES ↗ LATEST STABLE

Braintrust: Connects external MCP servers to Loop so it can use their tools and data alongside Braintrust traces, with per-project API key or OAuth connections shared with Loop and scheduled automations (requires data plane v2.16.0+).

└──▷ **TRY IT**

Deduplicate SQL query results to find the unique combinations of model and status across your traces.

```
$ SELECT DISTINCT model, status FROM traces WHERE project_id =
'<project_id>'
```

Enables Python workflow evaluations (SDK v0.39.0+) to run tasks and scorers through provider batch APIs, with support for polling or webhook-based resumption by any process after the provider completes work.

Maps Claude Code's native OpenTelemetry traces and events to LLM and tool spans with token metrics, cost, and conversation messages (requires data plane v2.16.0+).

Adds `bt` CLI (v0.19.3+) support for tracing Grok and Google Antigravity coding sessions as Braintrust traces, capturing prompts, model calls, and tool activity without interrupting the agent if tracing fails.

Adds Patterns, which runs Loop against your project on a schedule to detect recurring problems and show per-finding frequency with a chart for the selected time range (requires data plane v2.15.0+).

Adds model restrictions under AI provider governance to control which models each team can access, including a `Selected models` option for project providers on Braintrust-hosted deployments.

+2 more

└─▷ **BREAKING ON UPGRADE**

! On Braintrust-hosted deployments, project providers now override organization providers only for models they serve; organization providers can serve other models using their own credentials. If you rely on project providers to restrict model access or spending for all models, review organization provider model selections and permissions.

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI/LLM Security

NoLabs nono

Feedback ⓘ v0.79.0 NOTES ↗ CODE ↗ RANK

The nono runtime runs AI agents in isolated environments with policy-controlled filesystem, network, and credential access.

> nono v0.79.0 adds remote agent session launching and richer diagnostics, while docs detail a new environment-variable redaction system for diagnostics and audit logs; the release also removes legacy Claude hook cleanup as a breaking change.

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

01 Extra environment variable redaction for diagnostics NEW

95

Adds `diagnostics.redaction.extra_env_vars` config key to scrub additional environment variables — beyond well-known secrets and `environment.deny_vars` exact names — from diagnostics, session records, and the audit log; accepts exact names and case-insensitive globs like `ACME_*` and `*_SECRET`. Redaction targets are also auto-derived from `environment.deny_vars` exact-name entries, `command_policies.credentials` destinations, `network.custom_credentials` routes, and the `env_var` field of `credential_routes` entries (glob `deny_vars` entries are NOT auto-derived and must be repeated under `extra_env_vars`). Values merge additively across profile inheritance, and removing a secure default requires `[redaction].unsafe_redaction_overrides` in user config, with such removals recorded in the audit log's redaction-policy diff.

Prevent a deployment token and any org-scoped secrets from appearing in nono diagnostics or audit logs when they are passed through from the host shell.

yaml

```yaml
diagnostics:
  redaction:
    extra_env_vars:
      - "DEPLOY_TOKEN"
      - "ACME_*"
      - "*_SECRET"
```

Prevent sensitive env vars (e.g. AWS keys) from appearing in nono diagnostic output by listing them in your profile.

```yaml
diagnostics:
  redaction:
    extra_env_vars:
      - AWS_SECRET_ACCESS_KEY
      - GITHUB_TOKEN
```

— Names exact config key, related keys, glob syntax, and override mechanism v0.79.0

02 **Remote agent session launching from nono run** NEW 80

Adds `--remote`, `--agent`, `--workspace`, and `--detached` flags to `nono run` to launch and attach to remote agent sessions, e.g. running an agent detached in CI without keeping a terminal open.

Launch a remote Claude session and detach immediately — useful in CI or when you want an agent running without keeping a terminal open.

```
$ nono run --remote --agent claude --workspace my-project --detached
```

— Runnable command with exact flags shown v0.79.0

03 **Removal of legacy Claude hook cleanup** BREAKING 60

Removes legacy Claude hook cleanup code (`feat(cli)!: remove legacy Claude hook cleanup`); profiles that relied on the automatic cleanup of old Claude hooks will no longer have those hooks removed on startup.

— Names the removed mechanism and its effect on profiles v0.79.0

THINNER COVERAGE BELOW

04 **Proxy denial visibility in diagnostic footer** NEW 45

Renders proxy network denials in the diagnostic footer so blocked outbound calls are visible without manual log inspection.

— Describes behavior but no command or UI path given v0.79.0

05 **Standalone Snap runtime policy group** NEW 45

Adds a standalone Snap runtime policy group so Snap-packaged tools get correct sandbox rules without per-profile tuning.

— Names the feature but no config example given v0.79.0

└──▷ BREAKING ON UPGRADE

! Legacy Claude hook cleanup code is removed (`feat(cli)!: remove legacy Claude hook cleanup`); profiles that relied on the automatic cleanup of old Claude hooks will no longer have those hooks removed on startup.

v0.79.0 NOTES ↗ CODE ↗ LATEST STABLE

nono v0.79.0 adds remote agent session launching, proxy denial diagnostics, Snap policy group, and `diagnostics.redaction.extra_env_vars`

└──▷ **GET THIS VERSION**

```
$ git clone --branch v0.79.0 https://github.com/nolabs-ai/nono.git
# already have the repo? check out this version:
$ git checkout v0.79.0
```

└──▷ **TRY IT**

Launch a remote Claude session and detach immediately — useful in CI or when you want an agent running without keeping a terminal open.

```
$ nono run --remote --agent claude --workspace my-project --detached
```

Prevent sensitive env vars (e.g. AWS keys) from appearing in nono diagnostic output by listing them in your profile.

```yaml
diagnostics:
  redaction:
    extra_env_vars:
      - AWS_SECRET_ACCESS_KEY
      - GITHUB_TOKEN
```

Adds `--remote` , `--agent` , `--workspace` , and `--detached` flags to `nono run` to launch and attach to remote agent sessions.

Adds `diagnostics.redaction.extra_env_vars` profile key to control which additional environment variables are redacted from diagnostic output.

Renders proxy network denials in the diagnostic footer so blocked outbound calls are visible without manual log inspection.

Adds a standalone Snap runtime policy group so Snap-packaged tools get correct sandbox rules without per-profile tuning.

└──▷ **BREAKING ON UPGRADE**

! Legacy Claude hook cleanup code is removed (`feat(cli)!: remove legacy Claude hook cleanup`); profiles that relied on the automatic cleanup of old Claude hooks will no longer have those hooks removed on startup.

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Mistral AI

Feedback ⓘ snapshot-20260930 NOTES ↗ RANK

Learn how to deploy and use Mistral AI's Large Language Models with our comprehensive documentation, guides, and tutorials.

> Mistral shipped OCR 4.1 to general availability with block-level confidence scores, brought Z.ai GLM 5.3 to GA, and deprecated three older models with fixed retirement dates.

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

01 **Model deprecations and retirements** DEPRECATED 75

OCR 4.0 (`mistral-ocr-4-0`) is deprecated and retires on September 30, 2026, with callers pinned to it needing to migrate to `mistral-ocr-4-1` . Leanstral 1.5 (`labs-leanstral-1-5`) is deprecated and retires on September 30, 2026. Z.ai GLM 5.2 (`zai-glm-5-2`) is deprecated and retires on October 31, 2026, with callers required to migrate to `zai-glm-5-3` .

— Names retirement dates and migration targets for each model snapshot-20260930

THINNER COVERAGE BELOW

02 **OCR 4.1 general availability** NEW 35

OCR 4.1 (`mistral-ocr-4-1`) is now Generally Available, adding block-level confidence scores to OCR output.

— Only summary-level mention, no usage detail snapshot-20260930

03 **Z.ai GLM 5.3 general availability** NEW 30

Z.ai GLM 5.3 (`zai-glm-5-3`) is now Generally Available.

— Bare GA announcement, no further mechanism snapshot-20260930

└──▷ **BREAKING ON UPGRADE**

! OCR 4.0 (`mistral-ocr-4-0`) is deprecated and retires on September 30, 2026; callers pinned to it must migrate to `mistral-ocr-4-1` .

! Leanstral 1.5 (`labs-leanstral-1-5`) is deprecated and retires on September 30, 2026.

! Z.ai GLM 5.2 (`zai-glm-5-2`) is deprecated and retires on October 31, 2026; callers must migrate to `zai-glm-5-3` .

> snapshot-20260930 NOTES ↗ LATEST STABLE
>
> Mistral snapshot-20260930 brings OCR 4.1 GA with block-level confidence scores, Z.ai GLM 5.3 GA, and deprecations for OCR 4.0 and Leanstral 1.5.

Z.ai GLM 5.3 (`zai-glm-5-3`) is now Generally Available.

└─▷ **BREAKING ON UPGRADE**

- ! OCR 4.0 (`mistral-ocr-4-0`) is deprecated and retires on September 30, 2026; callers pinned to it must migrate to `mistral-ocr-4-1` .
- ! Leanstral 1.5 (`labs-leanstral-1-5`) is deprecated and retires on September 30, 2026.
- ! Z.ai GLM 5.2 (`zai-glm-5-2`) is deprecated and retires on October 31, 2026; callers must migrate to `zai-glm-5-3` .

WAS THIS USEFUL? 👍 👎

OpenAI GPT-6.1 Sol

Feedback ⓘ released · Sep 30, 2026

MODEL ↗

GPT-6.1 Sol from OpenAI was released on 29 Sep 2026. 1050K-token context; $2.00 in / $10.00 out per million tokens (OpenRouter).

WAS THIS USEFUL? 👍 👎

OpenAI GPT-6.1 Sol Pro

Feedback ⓘ

released · Sep 30, 2026 MODEL ↗

GPT-6.1 Sol Pro from OpenAI was released on 29 Sep 2026. 1050K-token context; $2.00 in / $10.00 out per million tokens (OpenRouter).

WAS THIS USEFUL? 👍 👎

OpenAI GPT-5.6 Sol Pro

Feedback ⓘ

price increase · Sep 29, 2026 MODEL ↗

GPT-5.6 Sol Pro is a multimodal model from OpenAI.

Price increase on OpenRouter (openai/gpt-5.6-sol-pro): $2.00 → $4.00 in, $10.00 → $20.00 out per million tokens (+100%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

xAI Grok 4.7

Feedback ⓘ price increase · Sep 29, 2026 MODEL ↗

Grok 4.7 is a reasoning model from xAI.

Price increase on OpenRouter (x-ai/grok-4.7): $1.60 → $2.00 in, $4.80 → $6.00 out per million tokens (+25%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

OpenAI gpt-oss-120b

Feedback ⓘ price cut · Sep 29, 2026

MODEL ↗

gpt-oss-120b is a reasoning model from OpenAI.

Price cut on OpenRouter (openai/gpt-oss-120b): $0.15 → $0.037 in, $0.60 → $0.17 out per million tokens (-75%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

◆ Open-Weight Models

Cohere Models

Feedback ⓘ September 30, 2026 NOTES ↗ RANK

Cohere develops Command, Aya, Embed, Rerank, transcription, and document-understanding models.

> Cohere launched Embed 5, a new embedding model family with two specialized variants sharing one embedding space, multimodal and multilingual support, flexible Matryoshka dimensions, and a 128k context window.

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

01 Matryoshka embeddings with flexible dimensions NEW

80

Supports Matryoshka embeddings in dimensions `[256, 512, 768, 1024, 1536, 2048]` with `float`, `int8`, and `binary` output types.

— Names exact dimension values and output types, a concrete configuration.

September 30, 2026

02 Embed 5 model variants: pro and fast NEW

70

Adds `embed-v5.0-pro`, optimized for highest retrieval quality, particularly for offline indexing and quality-critical retrieval, and `embed-v5.0-fast`, optimized for low latency and high throughput for interactive search, agent loops, and high-volume query traffic. The two variants share one embedding space, so a corpus indexed with one model can be queried with the other.

— Names both model ids and their intended use cases and shared-space mechanism.

September 30, 2026

THINNER COVERAGE BELOW

03 Embed 5 availability across platforms NEW

50

Available through the Embed API, Microsoft Foundry, Amazon SageMaker, and Model Vault for single-tenant deployment.

— Names concrete platforms a reader could deploy through today. September 30, 2026

04 Multimodal embedding inputs NEW

45

Embed 5 supports text, images, and mixed text-and-image inputs (e.g. PDF pages) encoded into a single vector.

— States capability but no usage detail or limits. September 30, 2026

05 Extended context window and language coverage IMPROVED

Extends context window to 128k tokens and supports over 100 languages for multilingual retrieval.

— Gives numbers but no mechanism or usage path. September 30, 2026

September 30, 2026 NOTES ↗ LATEST STABLE

Cohere Embed 5 launches with 128k context, multimodal inputs, Matryoshka embeddings, and two model variants sharing one embedding space.

Adds `embed-v5.0-pro` model variant optimized for highest retrieval quality, particularly for offline indexing and quality-critical retrieval.

Adds `embed-v5.0-fast` model variant optimized for low latency and high throughput, for interactive search, agent loops, and high-volume query traffic.

Shared embedding space between `embed-v5.0-pro` and `embed-v5.0-fast` allows a corpus indexed with one model to be queried with the other.

Supports multimodal inputs — text, images, and mixed text-and-image inputs (e.g. PDF pages) — in a single vector.

Supports Matryoshka embeddings in dimensions `[256, 512, 768, 1024, 1536, 2048]` with `float`, `int8`, and `binary` output types.

+3 more

WAS THIS USEFUL? 👍 👎

DeepSeek V3.1 Terminus

Feedback ⓘ

price increase · Sep 29, 2026 MODEL ↗

DeepSeek V3.1 Terminus is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v3.1-terminus): $0.27 → $0.30 in, $1.00 → $1.00 out per million tokens (+11%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

DeepSeek V4 Flash 0731 Feedback ⓘ

price increase · Sep 29, 2026 MODEL ↗

DeepSeek V4 Flash 0731 is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4-flash): $0.047 → $0.14 in, $0.0941 → $0.28 out per million tokens (+198%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

DeepSeek V4 Pro 0813 Feedback ⓘ

price increase · Sep 29, 2026 MODEL ↗

DeepSeek V4 Pro 0813 is a chat model from DeepSeek.

Price increase on OpenRouter (deepseek/deepseek-v4-pro): $0.6116 → $0.9553 in, $1.2232 → $1.9105 out per million tokens (+56%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

Z.ai GLM 5.1 Feedback ⓘ price increase · Sep 29, 2026 MODEL ↗

GLM 5.1 is a chat model from Z.ai.

Price increase on OpenRouter (z-ai/glm-5.1): $0.9646 → $1.40 in, $3.0316 → $4.40 out per million tokens (+45%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

MiniMax M2.7

Feedback ⓘ price cut · Sep 29, 2026 MODEL ↗

MiniMax M2.7 is a reasoning model from MiniMax.

Price cut on OpenRouter (minimax/minimax-m2.7): $0.30 → $0.21 in, $1.20 → $0.84 out per million tokens (-30%), first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

Alibaba Qwen3 30B A3B

Feedback ⓘ price cut · Sep 29, 2026

MODEL ↗ STABLE

Qwen3 30B A3B is a chat model from Alibaba.

Price cut on OpenRouter (qwen/qwen3-30b-a3b-instruct-2507): $0.10 → $0.0481 in, $0.30 → $0.193 out per million tokens (-52%), first seen 29 Sep 2026.

▾ 2 more releases in this issue · 2026-09-29

> price increase · Sep 29, 2026 MODEL ↗ STABLE
>
> Price increase on OpenRouter (qwen/qwen3-30b-a3b): $0.12 → $0.13 in, $0.50 → $0.52 out per million tokens (+8%), first seen 29 Sep 2026.

> retirement scheduled · Sep 29, 2026 MODEL ↗ STABLE
>
> OpenRouter lists a retirement date of 9 Oct 2026 for qwen/qwen3-30b-a3b, first seen 29 Sep 2026.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · September 30, 2026

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

▷ Subscribe Feedback on this issue