

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED OCTOBER 22, 2026 · EVERY DAY

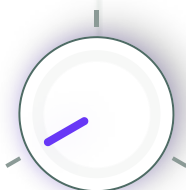
EDITIONS **tail** | grep | head | ls | diff | uniq | perf

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

[▶ // HOW THIS ISSUE IS MADE](#)[SHOW](#)

VIEW

ISSUE VIEW

scroll**SCROLL**EXPAND
IN
PLACEPREVIEW
PANE

INFORMATION DENSITY

everything

BRIEFED

EVERYTHING

Every tool in the issue, in full.

[all 69 changes](#)Do you prefer **scroll**?[▶ // CONTENTS](#) [SHOW](#)

53 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 today. OpenAI cut agent-loop latency with an ultrafast tier and pre-warmed sandboxes, and now bills boxes per turn rather than per idle minute. Composio lets you

block a tool call before it runs, using natural-language conditions. Anyscale's scheduler cohorts let idle GPU quota flow between queues instead of sitting reserved.

DEPLOY

Stop paying for agent sandboxes that are sitting idle between turns

OpenAI pre-warms Agents API sandboxes and moves them to turn-based billing, so the cold-start pause on the first tool call goes away and you are not charged for the wall-clock gap while a user reads the last answer. Anyscale's scheduler cohorts let resource queues share and preempt unused GPU/CPU quota, which is the usual fix for a cluster where one team's reserved capacity sits dark while another team queues.

[OpenAI](#) · [Anyscale](#)

GOVERN

Veto an agent's tool call before it executes, without writing a policy DSL

Composio's session hooks intercept a call before execution or before the result reaches the model, with conditions written in natural language and scoped per toolkit — the check lands at the boundary instead of in a prompt the model can be talked out of. ToolHive hardens the other end: TLS and SPIFFE on its auth server, plus limits and expiry on Dynamic Client Registration clients, so self-registered MCP clients do not accumulate permanent credentials.

[Composio](#) · [ToolHive](#)

BUILD

Get a calibrated probability back from a classifier instead of parsing free text

Shisa DE-2 returns calibrated probabilities for yes/no and multiple-choice questions, and StartLux-Decision v1.1 adds per-choice uncertainty with a tunable confidence gate. A number you can threshold means routing and human-escalation logic becomes a comparison rather than a regex over whatever phrasing the model chose this time.

[Shisa](#) · [StartLux](#)

Replit Agent

Feedback



changelog · Oct 22, 2026

CHANGELOG >

RANK

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

Replit Agent's window centered on infrastructure flexibility and AI model choice: a new Bring Your Own Cloud offering with three deployment tiers, Claude Haiku 5.5 added to Power Mode, and a new Solutions Partners directory.

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

01 Bring Your Own Cloud deployment tiers NEW 73

Replit introduces a Bring Your Own Cloud (BYOC) offering with three tiers: Full BYOC (private preview) to run the full Replit platform in your own cloud tenant, BYOC for Published Apps (early access) to host published apps in your cloud tenant, and Image Exports (early access) to export and self-host Replit-built apps.

Evaluate BYOC for Published Apps when your enterprise requires apps to run in your own cloud tenant rather than Replit-managed infrastructure.

📌 Navigate to the BYOC documentation at docs.replit.com, then contact Replit to request early access to BYOC for Published Apps and follow the provisioning steps to host published Replit apps in your cloud tenant.

– Names all three tiers and access status but lacks setup mechanics

changelog-20261022-2b1ee380

THINNER COVERAGE BELOW

02 Claude Haiku 5.5 in Power Mode model selector NEW 58

Adds Claude Haiku 5.5 as a selectable model in the Power Mode model selector, available to Core and Pro plan builders.

Select Claude Haiku 5.5 for faster, lower-latency AI assistance in Power Mode on a Core or Pro plan.

📌 In the Replit editor, open Power Mode, click the model selector dropdown, and choose Claude Haiku 5.5.

– Clear UI path to select the model, but no detail on performance or mechanism

changelog-20261022-2b1ee380

Launches a directory to browse and connect with agencies, consultants, and systems integrators that build with Replit, with an application path for firms to join the program.

– Describes purpose but no concrete navigation or named surface

changelog-20261022-2b1ee380

WAS THIS USEFUL?



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

Warp adds a rebindable keyboard shortcut for toggling Markdown rendering, recognizes Muse Code as a first-class CLI agent, and extends orchestrator and cloud-agent run metadata.

WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01

Markdown rendered/raw toggle shortcut

NEW

75

Warp adds a rebindable keyboard shortcut — `⌘+E` on macOS, `Ctrl+E` on Windows/Linux — to toggle open Markdown files between rendered and raw view. To rebind it, go to Settings > Keyboard Shortcuts and search for `Toggle Markdown view`.

Toggle a Markdown file between its rendered preview and raw source without leaving the keyboard.

Open any .md file in Warp, then press `Cmd+E` (macOS) or `Ctrl+E` (Windows/Linux) to switch between rendered and raw view. To rebind: go to Settings > Keyboard Shortcuts and search for `Toggle Markdown view`.

– Exact shortcut and rebind path given `changelog-20261022-2d43208a`

THINNER COVERAGE BELOW

02

Muse Code recognized as CLI agent

NEW

40

Warp now recognizes Muse Code as a first-class CLI agent, with branded session icons and rich-input support.

– Names the agent but no usage steps `changelog-20261022-2d43208a`

03

Orchestrator batch and child run metadata

NEW

35

Orchestrators can now attach batch and per-child run metadata for grouping and setup workflows.

– Describes capability without concrete fields or commands `changelog-20261022-2d43208a`

04

Cloud agent workspace handoff checkpoints

IMPROVED

20

Cloud agents periodically save workspace handoff checkpoints during a run.

– Brief mechanism description, no specifics given `changelog-20261022-2d43208a`

ALSO FROM THESE RELEASES

.

WAS THIS USEFUL?



Vercel v0

Feedback



changelog · Oct 22, 2026

CHANGELOG

RANK

The v0 generative UI platform builds web applications from natural-language prompts.

v0 shipped Slack MCP integration for pulling context into chats, a ChatGPT Plus/Pro billing option that offsets v0 credit usage, and several iOS app and Code tab refinements.

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

01 Download action for large unpreviewable files

IMPROVED

45

On the Code tab, files too large to preview now show a Download action in the file header instead of a blank preview.

– Names exact location but no further detail on size thresholds.

changelog-20261022-d0b2dbc6

02 iOS app pickers and preview additions

NEW

40

The v0 iOS app gains a searchable skills picker, a design system picker, and Quick Look video attachment previews.

– Lists what was added but no usage steps or mechanism.

changelog-20261022-d0b2dbc6

WAS THIS USEFUL?



yetone magpie

Feedback



v0.1.1321

CODE ↗

RANK

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

Magpie's latest releases sharpen the redact engine with case-insensitive masking and smarter ID preservation, while also adding dsh session titles in Context, MiniMax M3.1-Flash effort configuration, and a Qoder mover version gate.

↳ WHAT SHIPPED · 5 FEATURES most completely described first ? what's the number?

01 Case-insensitive masked-word matching in redact engine

IMPROVED

70

Adds a `WordsAnyCase` option to the redact engine, toggled via the `Masked words` switch in settings, so masked words match regardless of case (e.g. `nightjar`, `NIGHTJAR`, and `Nightjar` all become placeholders) while each is restored to its original casing.

– Names the option, UI switch, and behavior with examples. v0.1.1319

02 Backend ID preservation in redact engine

IMPROVED

65

The redact engine now preserves backend IDs (UUIDs, 32+ hex-digit hashes) as written everywhere except fields explicitly named for credentials (e.g. `password=`, `api_key:`, `refresh_token:`), preventing false-positive masking of row, request, session, and commit IDs.

– Clear scope and before/after but no user-facing action. v0.1.1319

THINNER COVERAGE BELOW

03 Qoder mover minimum version gating

IMPROVED

55

Sets the Qoder mover minimum version to `0.2.28`, which queues requests through Qoder's own queue before forwarding.

– Names version and mechanism but no reader action. v0.1.1321

04 Session titles shown in Context page

IMPROVED

50

Context page now keys dsh gateway sessions by their titles rather than the internal `magpie-a...` hash, matching the display already used on the Sessions page.

– Describes the UI change but it's automatic, no action needed. v0.1.1321

05 MiniMax M3.1-Flash effort configuration

NEW

35

Adds MiniMax M3.1-Flash effort configuration via `output_config` .

– Only named in summary, no mechanism or example given. `v0.1.1321`

WAS THIS USEFUL?



AGENT

Composio

[Feedback](#)[@composio/slim@0.27.0](#)[NOTES](#)[RANK](#)

Composio connects AI agents to external tools and services with managed authentication and integrations.

Composio shipped session hooks that let developers intercept and block AI agent tool calls—before they execute or before results reach the agent—based on natural-language conditions scoped to specific toolkits.

↳ WHAT SHIPPED • 1 FEATURE most completely described first ? what's the number?

01 Session hooks to block tool calls before execution

BREAKING

95

The `hooks` parameter on `create()` and `update()` registers pre-execution or pre-result interceptors; a `before_execute` hook with a `condition_check` handler evaluates a natural-language `condition` and can `action: 'block'` a tool call before the agent sees it or its output. Hooks can be scoped with `appliesTo` (e.g. `{ toolkits: ['gmail'] }`) to target specific toolkits, and the response shape of `session.execute()`, `session.executeMeta()`, `session.proxyExecute()`, and `ctx.proxyExecute()` now includes `resultType: 'completed'` or `resultType: 'blocked'`, with blocked calls carrying `error`, `blockReason`, and `hookResults` instead of a raw HTTP response. On `update()`, passing `null` for `hooks` clears every registered hook, while a new array replaces the whole list; `hooks` cannot be combined with `experimental.sessionConfigId` on either `create()` or `update()`. This is a breaking change for TypeScript callers of `proxyExecute()` that read `status` directly, since a blocked request has no HTTP response and `status` is absent when `resultType` is `'blocked'`.

Block outbound Gmail messages to addresses outside your domain before the agent ever sends them.

```

const session = await composio.create('user_123', {
  hooks: [
    {
      type: 'before_execute',
      id: 'no_external_email',
      name: 'No external email',
      appliesTo: { toolkits: ['gmail'] },
      handler: {
        type: 'condition_check',
        condition: 'Sending email to anyone outside example.org',
        action: 'block',
      },
    },
  ],
});

```

After a tool call, check whether a hook blocked it and surface the reason to the user instead of silently failing.

```

const result = await session.execute(toolCall);
if (result.resultType === 'blocked') {
  console.error('Blocked by hook:', result.blockReason,
    result.error);
}

```

Clear all hooks from an existing session during a live update.

```

await composio.update('user_123', { hooks: null });

```

– Full mechanism, exact API/config names, and runnable code examples.

[@composio/slim@0.27.0](#)

[@composio/core@0.27.0](#)

⚠️ BREAKING ON UPGRADE

- ! TypeScript callers of `session.proxyExecute()` and `ctx.proxyExecute()` that read `status` directly must now check `resultType` first — a blocked request carries no HTTP response, so `status` is absent when `resultType` is `'blocked'`.
- ! `hooks` cannot be combined with `experimental.sessionConfigId` on `create()` or `update()`; mixing them is an error.
- ! TypeScript code reading `status` directly on the result of `session.proxyExecute()` or `ctx.proxyExecute()` will break for blocked requests — callers must check `resultType`

before accessing `status`.

WAS THIS USEFUL?



Cloudflare Agents

[Feedback](#)[agents@0.30.0](#)[NOTES >](#)[RANK](#)

The Cloudflare Agents SDK builds stateful agents on Workers and Durable Objects, with persistent state, scheduling, WebSockets, and MCP server support.

Cloudflare Agents shipped Pi 1.4 support for PiHarness, adding hibernation during idle Pi tasks, recovery for crash-looping tasks, and automatic recovery from storage errors — alongside breaking changes from the Pi 1.4 upgrade itself.

↳ [WHAT SHIPPED](#) · [4 FEATURES](#) most completely described first [? what's the number?](#)

01 Durable Object hibernation via onWake in PiHarness

NEW

85

PiHarness factories can now pass an **onWake** parameter through to `Harness.open(storage, { ...options, onWake }, context)`, letting the Durable Object hibernate while Pi's tasks sleep instead of staying resident in memory.

Enable Durable Object hibernation while Pi tasks sleep by wiring **onWake** through your harness factory.

```
typescript

import { PiHarness } from 'agents/harness/pi';
import { Harness } from '@earendil-works/pi-durable';
import { Agent } from 'agents';

export class Assistant extends Agent<Env> {
  registry = createRegistry();
  harness = new PiHarness({
    harness: ({ storage, context, onWake }) =>
      Harness.open(storage, { models, registry: this.registry,
onWake }, context)
  });
}
```

— Names the exact API and includes a runnable wiring example.

[agents@0.30.0](#)

02 Parked task status and retry recovery in PiHarness

NEW

85

`PiHarness.pending()` now reports a **parked** status for tasks stopped by Pi's restart limit, and a new `retry()` method re-runs them so crash-looping operations don't silently block.

Surface and re-run tasks that Pi's restart limit has parked, so crash-looping operations do not silently block.

```
typescript

// Check for parked tasks
const ops = await agent.harness.pending();
const parked = ops.filter(op => op.status === 'parked');

// Re-run each parked task
for (const op of parked) {
  await agent.harness.retry(op.id);
}
```

— Names the status and method, shown with a runnable check-and-retry example.

agents@0.30.0

03 Pi 1.4 upgrade and breaking changes in agents/harness/pi

85

BREAKING

`agents/harness/pi` now supports Pi `~1.4.0` (`@earendil-works/pi-durable`, `pi-ai`, `pi-agent-core`), which migrates SQLite session storage to a new schema that Pi 1.3 cannot open, so deployments cannot roll back past this release. The upgrade removes `TaskRuntime.sleep()`, introduces a `sleeping` task status and a restart limit that parks a crash-looping task after 5 restarts (code switching on `PiPendingOperation.status` must now handle the new `parked` value), adds `@earendil-works/chord` as a required new peer dependency alongside `pi-durable` and `pi-ai`, and requires that the Pi registry no longer be created at module scope — move it into your class as a field or inside the factory, or evicted objects will keep their harness in memory indefinitely.

— Names every package, API, status, and migration step with clear remediation.

agents@0.30.0

THINNER COVERAGE BELOW

04 Automatic recovery from PiHarness storage errors

IMPROVED

40

`PiHarness` now reopens Pi automatically after a storage error instead of failing every subsequent call until the Durable Object restarts.

— Explains the behavior change but no example or config surface given.

agents@0.30.0

↳ BREAKING ON UPGRADE

- ! Pi 1.4 migrates SQLite session storage to a schema Pi 1.3 cannot open — a deployment cannot be rolled back past this release.

- ! `TaskRuntime.sleep()` is removed in Pi 1.4 (see pi-durable 1.4.0 changelog).
- ! Pi 1.4 introduces a `sleeping` task status and a restart limit that parks a crash-looping task after 5 restarts — code that switches on `PiPendingOperation.status` must handle the new `parked` value.
- ! The Pi registry must no longer be created at module scope; move it into your class as a field or inside the factory, or evicted objects will keep their harness in memory indefinitely.

WAS THIS USEFUL?



DEPLOY

Nebius AI Cloud

Feedback



changelog · Oct 22, 2026

CHANGELOG ↗

Nebius AI Cloud provides GPU compute, managed Kubernetes, distributed training infrastructure, and serverless AI workloads.

Adds browser-based Cloud Shell (preview), Kubernetes maintenance-deferral Pod labels, and private-service webhook ingress via Nebius Tunnels.

Adds Cloud Shell, a browser-based terminal in the web console with the Nebius AI CLI, Terraform, Git, and Python preinstalled and pre-authenticated; sessions last up to one hour.

Assign the `nebius.com/defer-maintenance-node-drain: "true"` label to CPU Pods to defer Managed Kubernetes node drain until one hour before the SLA deadline.

UPDATED SUPPORTED PLATFORM

Enables receiving webhooks from providers such as GitHub, Stripe, and Telegram in a private service (no public IP, no open inbound ports) via Nebius Tunnels.

Raises Nebius Tunnels traffic limits to 120 Mbps per tunnel, 600 Mbps per project, and 8 Gbps per region (inbound and outbound combined), with client connection limits now applied per new connections per second.

Sets the Enhanced Throughput storage quota for total object size to 10 TiB in all regions.

UPDATED SUPPORTED PLATFORM

Serverless AI regional availability

⚠️ BREAKING ON UPGRADE

! Devlabs now require a container disk of at least 50 GiB for GPU workloads or 20 GiB for CPU-only workloads; workloads that do not meet this minimum will fail to launch.

WAS THIS USEFUL?



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

Anyscale 0.26.111 introduces scheduler cohorts that let resource queues share and preempt unused GPU/CPU quota, alongside cloud-scoped compute-config archiving, a force-clean option for duplicate skill hooks, and several CLI/SDK reliability fixes.

← WHAT SHIPPED · 7 FEATURES most completely described first ? what's the number?

01 Scheduler cohort quota-sharing for resource queues

NEW

90

Adds `cohort_name`, `lending_limit`, `borrowing_limit`, `reclaim_within_cohort`, and `borrow_within_cohort` fields to scheduler config resource queues, enabling cohorts that share unused quota across queues according to preemption policy (`never`, `lower_priority`, `any`).

Define a scheduler cohort so two resource queues can lend and borrow unused GPU quota from each other, with preemption controls.

```
yaml

cohort_name: gpu-team-cohort
nominal_quota:
  GPU: 8
lending_limit:
  GPU: 4
borrowing_limit:
  GPU: 4
preemption:
  reclaim_within_cohort: lower_priority
  borrow_within_cohort: lower_priority
```

– Config fields and example given, but no CLI command shown

[changelog-20261022-096b2a2c](#)

02 Force-clean duplicate skill hooks

NEW

85

Adds `anyscale skills update --force` to immediately remove duplicate Anyscale hook entries that accumulated when agents such as Claude Code rewrote their `settings.json`.

Force-remove duplicate Anyscale skill hooks that accumulated after an agent like Claude Code rewrote its settings.

```
$ anyscale skills update --force
```

– Exact runnable command and cause explained changelog-20261022-096b2a2c

03 Cloud-scoped compute-config archive NEW 83

Adds `--cloud` flag to `anyscale compute-config archive` and a `cloud` parameter to `anyscale.compute_config.archive()` to scope archive lookups to a single cloud when multiple clouds share the same compute-config name.

Archive a compute config by name scoped to one cloud, avoiding ambiguity when the same name exists in multiple clouds.

```
$ anyscale compute-config archive my-gpu-config --cloud my-aws-cloud
```

– Runnable command and exact flag given changelog-20261022-096b2a2c

04 Gzip log decompression for job logs IMPROVED 70

Improves `anyscale job logs`, `anyscale job submit --wait`, `anyscale.job.get_logs()`, and `anyscale.job.wait(follow=True)` to decompress gzip-compressed log files returned without a `Content-Encoding` header, as some S3-compatible storage does.

– Named commands but no usage example given changelog-20261022-096b2a2c

05 Changed return type for `terminate_system_cluster()` BREAKING 60

`anyscale.cloud.terminate_system_cluster()` now returns a list of terminated system-cluster IDs instead of a single ID, and returns an empty list (instead of failing) when no system cluster is running.

– Breaking change noted but no migration code shown changelog-20261022-096b2a2c

THINNER COVERAGE BELOW

06 Clearer CLI/API error messages IMPROVED 45

Server errors now show the message and a trace ID (`Internal Server Error (trace ID: <trace-id>)`), and validation errors show the field message instead of a raw field-error list.

– Behavior described but no actionable surface named changelog-20261022-096b2a2c

07 New Importing state in listing commands IMPROVED 35

Adds **Importing** to the states accepted by `--state` on the relevant listing commands.

– Thin description, commands not named `changelog-20261022-096b2a2c`

↳ **BREAKING ON UPGRADE**

! `anyscale.cloud.terminate_system_cluster()` now returns a list of IDs instead of a single string ID — code that treats the return value as a string will break.

WAS THIS USEFUL?



◆ **Agentic Infrastructure**

Snowflake Cortex AI

Feedback



changelog · Oct 22, 2026

CHANGELOG ↗

RANK

Snowflake Cortex AI provides managed language-model functions, search, and agents that work with data governed in Snowflake.

Snowflake Cortex AI added retrieval-augmented search, tool-calling in its AI Gateway with wider regional availability, and three new preview language models this cycle.

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

01 **Cortex AI Gateway: tool-calling and new regions** **NEW**

50

Adds **Tools for Cortex AI Gateway** (Preview), enabling tool-calling through the Cortex AI Gateway, and extends the **Cortex AI Gateway** to Azure, Google Cloud, and AWS New Zealand regions.

– Names surfaces and regions but no config or endpoint detail `changelog-20261022-30e4947e`

02 **New preview models: Grok 4.7, GLM 5.3, Gemini 4 Argon** **NEW**

48

Adds **Grok 4.7** and **GLM 5.3** models (Preview) in Cortex AI, and the **Gemini 4 Argon** model (Preview) in AI Functions.

– Names exact models and surfaces but no usage guidance `changelog-20261022-30e4947e`

WAS THIS USEFUL?



DATA

Pinecone

[Feedback](#)[changelog · Oct 22, 2026](#)[CHANGELOG >](#)[RANK](#)

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

Pinecone's single release in this window adds Admin API cost/usage exports, on-demand index capacity in BYOC, a new cross-region restore fee, and retires Pinecone Local.

↳ **WHAT SHIPPED · 4 FEATURES** most completely described first [? what's the number?](#)

01 Admin API usage and cost export NEW 70

Enterprise organizations can now export usage and cost data via the **Admin API**, covering any period in the last 365 days, with output delivered as a zip archive of FOCUS v1.4 CSV files.

– Names API and format but no exact endpoint path.

[changelog-20261022-1bfda8a1](#)

02 On-demand indexes in BYOC NEW 60

BYOC environments on AWS, GCP, and Azure now support on-demand indexes, letting you choose between on-demand or dedicated read nodes as the capacity mode when creating an index.

– Clear capability and providers named, no exact command shown.

[changelog-20261022-1bfda8a1](#)

03 Pinecone Local end of life BREAKING 60

Pinecone Local, the in-memory Docker emulator, has reached end of life and is no longer available; applications relying on it must migrate to a separate Pinecone account index or mock the Pinecone API.

– States impact and two migration paths, no step-by-step guide.

[changelog-20261022-1bfda8a1](#)

THINNER COVERAGE BELOW

04 Cross-region backup restore fee NEW 50

Restoring a backup into a different region now incurs a cross-region transfer fee, charged per GB and billed as **Serverless Cross-Region Transfer** on your invoice.

– Exact billing line named but no action to take.

[changelog-20261022-1bfda8a1](#)

↳ **BREAKING ON UPGRADE**

! Pinecone Local (the in-memory Docker emulator) has reached end of life and is no longer available; applications relying on it must migrate to a separate Pinecone account index or mock the Pinecone API.

WAS THIS USEFUL?



GOVERN

◆ AI/LLM Security

cplt

Feedback



2026.10.22-115127-f0950ea

NOTES ↗

RANK

Sandbox for AI coding agents. Runs Copilot CLI, Claude Code, OpenCode, Gemini CLI, Antigravity, Pi, goose or a plain shell inside a kernel-level sandbox, with git and gh guards and sandbox policy committed to the repository.

cplt added automatic authentication for GitHub Copilot CLI on Linux by reading the host's keyring login inside the sandbox.

↳ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 Linux keyring auth for Copilot CLI

NEW

45

cplt now reads GitHub Copilot's Linux keyring login from the host inside the sandbox, so Linux users no longer need to manually export a token to authenticate.

– Explains mechanism but no exact command or config shown

2026.10.22-115127-f0950ea

WAS THIS USEFUL?



ToolHive

Feedback



v0.56.0

NOTES

RANK

ToolHive runs MCP servers in isolated containers and enforces identity and access policies for local and Kubernetes deployments.

ToolHive v0.56.0 hardens its auth server with TLS and SPIFFE support, adds configurable limits and expiry for Dynamic Client Registration (DCR) clients, and tightens token session handling.

WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 TLS/SPIFFE listener for auth server NEW 57

Adds a TLS listener and SPIFFE middleware to the auth server; the TLS listener is also started in vMCP.

– Names TLS, SPIFFE, vMCP but no config flag given v0.56.0

02 DCR client cap and expiry controls NEW 57

Exposes DCR (Dynamic Client Registration) client cap and register rate limit as configurable fields in the CRD, and caps DCR clients while expiring unused ones after 24 hours.

– Mechanism and timing given, exact field names not provided v0.56.0

03 HTTP 401 for sessions missing upstream token IMPROVED 40

Returns HTTP 401 when a token session holds no upstream access token.

– Specific status code but no endpoint named v0.56.0

04 Skip OAuth consent for pre-authorized clients NEW 35

Allows embedders to skip the OAuth consent step for clients that are pre-authorized.

– Behavior described but no API or flag named v0.56.0

WAS THIS USEFUL?

AI MODELS

OpenAI

[Feedback](#)[changelog · Oct 22, 2026](#)[CHANGELOG](#)[RANK](#)

OpenAI provides AI models and APIs for building applications that generate and reason with text, images, and audio.

OpenAI's October update focuses on latency and cost in the Agents API — an ultrafast service tier, sandbox pre-warming, and turn-based sandbox billing — alongside a new prompt-to-voice endpoint, longer prompt caching, an Artifact SDK for document editing, and expanded GPT-Live voice and decision capabilities.

←→ WHAT SHIPPED · 8 FEATURES most completely described first ? what's the number?

01 Ultrafast service tier and sandbox pre-warming NEW

90

Adds `agent.service_tier` set to `ultrafast` to the Agents API (`v1/agents`), reducing inter-token latency with supported models, and adds sandbox pre-warming, letting OpenAI-hosted sandboxes be created and configured before a user finishes sending their message to cut session startup latency.

Reduce time-to-first-token in an Agents API session by enabling the ultrafast service tier on a supported model.

```
$ POST https://api.openai.com/v1/agents
Content-Type: application/json
Authorization: Bearer $OPENAI_API_KEY

{
  "model": "<supported-model>",
  "agent": {
    "service_tier": "ultrafast"
  }
}
```

— Names exact field, value and endpoint with a runnable call [changelog-20261022-9067d2cc](#)

02 Prompt-to-voice generation via API NEW

70

Adds `prompt-to-voice` via `v1/audio/voices` for all API customers: create a custom voice from a text description through the API or voice Playground, then reuse it in GPT-Live sessions.

— Names endpoint and reuse path but no sample request [changelog-20261022-9067d2cc](#)

03 Per-project service tier restrictions NEW 65

Enables per-project service tier restrictions in Platform settings, with organization-level controls that can restrict service tiers across all projects. Configure allowed tiers under Organization Settings > Service Tiers, then select permitted tiers for a given project in Platform Settings.

Restrict which service tiers a specific project can access to prevent accidental use of premium tiers.

 In the OpenAI Platform, go to Organization Settings > Service Tiers and configure allowed tiers at the org level, then go to Platform Settings for the target project and select the permitted service tiers.

– Exact UI navigation path to configure the control [changelog-20261022-9067d2cc](#)

04 12-hour prompt caching option NEW 60

Adds a 12-hour prompt caching option in the API, keeping shared instructions, tools, and context cached longer for lower cached-input pricing; 12-hour cache writes cost 2.5× the input token price.

– Gives pricing multiplier but no sample call [changelog-20261022-9067d2cc](#)

THINNER COVERAGE BELOW

05 Decisions integration in GPT-Live NEW 40

Adds native Decisions integration in GPT-Live, supporting direct delegation from Live or triggering decisions as input/output transcript deltas arrive.

– Describes mechanism but no config or endpoint named [changelog-20261022-9067d2cc](#)

06 Artifact SDK for document editing in sandboxes NEW 35

Adds Artifact SDK and skills for creating and editing spreadsheets, presentations, and Word documents to OpenAI-hosted sandboxes in the Agents API.

– Names SDK and formats but no usage mechanism [changelog-20261022-9067d2cc](#)

07 Nine new GPT-Live voices NEW 35

Adds nine new natural voices for GPT-Live: Nuri, Juni, and Flitz (German); Brise and Sillage (French); Nira (Hindi); Harema and Shida (Japanese); and Noeul (Korean).

– Names all voices but no usage instructions [changelog-20261022-9067d2cc](#)

08 Turn-based sandbox billing IMPROVED 35

Changes OpenAI-hosted sandbox billing to charge only during active turns, eliminating charges for idle time between turns.

WAS THIS USEFUL?



ScaleDown Extract

Feedback



released · Oct 20, 2026

MODEL ↗

Extract is a multimodal model from ScaleDown.

Extract from ScaleDown was released on 13 Oct 2026. 1M-token context.

WAS THIS USEFUL?



Elorian Glimpse

Feedback



released · Oct 20, 2026

MODEL ↗

Glimpse is a multimodal model from Elorian.

Glimpse from Elorian was released on 17 Oct 2026. 262K-token context; \$0.60 in / \$3.00 out per million tokens (OpenRouter).

WAS THIS USEFUL?



Inception Mercury 3

Feedback



released · Oct 20, 2026

MODEL ↗

Mercury 3 is a chat model from Inception.

Mercury 3 from Inception was released on 16 Oct 2026. 1M-token context; \$0.10 in / \$0.35 out per million tokens (OpenRouter).

Succeeds Mercury 2.5.

WAS THIS USEFUL?



Meta Muse Wonder

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

Muse Wonder is a reasoning model from Meta.

Muse Wonder from Meta was released on 12 Oct 2026. 1049K-token context; \$4.00 in / \$20.00 out per million tokens (OpenRouter).

WAS THIS USEFUL?



Unbiased Pareto 26.10

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)[BLOG ↗](#)

Pareto 26.10 is a multimodal model from Unbiased.

Unbiased: "Build with the stable Pareto 26.10 API: 1M context, \$0.80 input and \$3.20 output per million tokens. Explore pricing, weekly plans and preliminary benchmark..."

Unbiased says: "Pareto 26.10 API pricing: \$0.80 input, \$0.03 cached input and \$3.20 output per million tokens."

Unbiased says: "Pareto 26.10 \$1.56 Pareto 26.10: \$1.56 estimated API cost"

Unbiased says: "Pareto 26.10 \$1.56; Opus 5.5 \$8.80; Sol 6.1 \$4.40; Fable 5.1 \$21.00; Astra 6 \$24.00."

Pareto 26.10 from Unbiased was released on 15 Oct 2026. 1049K-token context; \$0.80 in / \$3.20 out per million tokens (OpenRouter).

Succeeds Pareto 26.9.

WAS THIS USEFUL?



Unbiased Pareto 26.9

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

Pareto 26.9 is a multimodal model from Unbiased.

Pareto 26.9 from Unbiased was released on 13 Oct 2026. 262K-token context; \$2.50 in / \$7.50 out per million tokens (OpenRouter).

WAS THIS USEFUL?



ScaleDown Summarize

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

Summarize is a multimodal model from ScaleDown.

Summarize from ScaleDown was released on 13 Oct 2026. 1M-token context.

WAS THIS USEFUL?



Unbiased

Feedback



Tools

NOTES ↗

RANK

Unbiased is a frontier answers at open-source prices. Zero data retention on designated API traffic (not Chat).

Unbiased shipped a Tokenizer Tax calculator that estimates Claude output costs across six models, processing input entirely client-side.

↳ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 Tokenizer Tax cost calculator NEW

60

A new tool lets users enter words-per-month or paste sample text to estimate output cost across six Claude models, using characters-per-output-token data from Verbosity Index Edition 2. Pasted text is processed entirely in-browser, keeping input data local.

– Describes mechanism and data source but no UI path or limits. Tools

WAS THIS USEFUL?

◆ Open-Weight Models

H2O.ai h2o-lightning-31b Feedback ⓘ

released · Oct 22, 2026 MODEL ↗

h2o-lightning-31b from H2O.ai was released on 20 Oct 2026. Open weights on Hugging Face.

WAS THIS USEFUL?

ConvAI Innovations Laya

[Feedback](#)[v0.4.13](#)[NOTES ↗](#)

RANK

Non-autoregressive System 1 decision engine. Typed choice, score and yes/no decisions over any text in a single forward pass, in 100+ languages, with a router that picks the right checkpoint per request.

Across eleven releases Laya broadened its SDK surface — new Rust (`laya-onnx`) and Go (`laya-go`) clients, a Java dispatch router, and `predict_long` / `PredictLong` long-context parity across Rust, .NET, MCP and Go — while adding answer-by-elimination tournament selection for large choice sets, Jev confidence reporting, and expanded training-pipeline evaluation tooling.

→ WHAT SHIPPED · 16 FEATURES most completely described first ? what's the number?

01 Tournament-based selection for large choice sets NEW

90

Adds the `LAYA_TOURNAMENT_GROUP` environment variable to enable answer-by-elimination tournament selection for large choice sets in the `serve` component, exposes core tournament selection as the `laya_tournament` MCP tool, extends large-choices narrowing to the `/v1/systemone/batch` endpoint, attributes tournament elimination and final-choice errors in evals reporting, and ports `predict_tournament` as `Shortlist.predictTournament` in the Java SDK.

— Names env var, endpoint, MCP tool and Java method across the flow

[v0.4.7](#)[v0.4.3](#)[v0.4.2](#)

02 Long-context prediction across SDKs NEW

85

Adds support for running inference on states that exceed the model's context window across SDKs: `predict_long` in the Rust SDK (matching Python SDK parity), `PredictLong` in the .NET client (ported from `ONNXAgent.predict_long`), `laya_predict_long` in the MCP interface, and `PredictLong` in `laya-go`.

— Names exact methods across four SDKs but no usage example

[v0.4.13](#)[v0.4.12](#)[v0.4.11](#)[v0.4.8](#)

03 New language SDKs: 'laya-go' and 'laya-onnx' NEW

83

Ships `laya-go`, a pure-Go tokenizer and language detection foundation with a Python parity lane, then adds client support for questions, sequences, and answer decoding, a `min_confidence` gate for `Predict` and `PredictBatch`, per-language temperatures, and checkpoint digest verification, and adds `laya-onnx`, a Rust SDK for running Laya inference from Rust.

– Traces laya-go from foundation to typed methods and config options Names the crate but no usage detail beyond existence

v0.4.9

v0.4.7

v0.4.4

v0.4.11

04 Training pipeline evaluation tooling NEW 78

Adds an `--eval` flag, abstention threshold fitting, and before/after `train_report.json` output to the training pipeline, a three-seed RLCD vs soft-CE default-loss comparison panel for training research, and prints the optimizer-update budget before a training run starts.

– Names the flag and output file but no exact command v0.4.1 v0.3.29

05 'LAYA_EXTRA_MODELS' checkpoint registration NEW 75

Adds the `LAYA_EXTRA_MODELS` environment variable to register extra checkpoints at serve time.

– Exact env var name with its effect at serve time v0.4.1

06 Java SDK dispatch router and language routing NEW 73

Adds `Router.predictBatch` to the Java SDK for batched router decisions, Java dispatch router lifecycle hooks with a new `AsyncHook` interface, and Java language routing with presets and an embedding shortlist.

– Names Java classes and interfaces but gives no code sample v0.4.2 v0.4.1

07 Agent and tool integrations NEW 70

Forwards `task` and `lang_guess` Router-only controls through the integrations layer, adds the `laya[agno]` integration for Agno agent routing and guardrails, and adds `LayaToolSelector` for tool and function routing.

– Names the integration package and selector, mechanism left thin v0.4.11

08 Jev confidence formula and 'x_jev_confidence' field NEW 70

Exports `jev_confidence`, the formula underlying the served `x_jev_confidence` field, for direct use in confidence calculations, and reports it as `x_jev_confidence` in laya-serve responses.

– Names the field and formula but no numeric example v0.4.3 v0.4.2

09 TypeScript SDK API additions IMPROVED 65

Ports `decideBatch` and exposes `answerConfidence` in the TypeScript SDK, and honours a question's `option_order` across encoding and decoding.

– Names three TS methods/fields without broader context v0.4.11 v0.4.7

10 Option stability hook and diagnostics IMPROVED 60

Adds `OptionStabilityHook` to the hooks system as a metamorphic reliability signal, then extends it with local option stability diagnostics in the TypeScript layer.

– Names the hook but the 'metamorphic' mechanism stays vague v0.4.13 v0.4.12

THINNER COVERAGE BELOW

11 Decode optimization in `'_decode_answers'` IMPROVED 52

Optimizes `logit` decoding and confidence calculation in `_decode_answers` for faster agent inference.

– Names the function but gives no benchmark or numbers v0.4.13

12 Cross-platform setup launchers NEW 51

Adds `start.bat` to launch the web UI on Windows, and a platform-detecting setup entry point with Linux setup support.

– Names the launcher files but no further detail v0.4.11 v0.4.8

13 Evals subsystem improvements NEW 50

Adds student-agreement comparison with repeated teacher decisions in the evals module, and evidence inspection over checkpoint and eval artifacts in the evals subsystem.

– Names the two additions but not how to invoke them v0.4.2 v0.3.29

14 Router warning for near-duplicate registered names IMPROVED 43

Adds a warning in the router when a registered name is one edit away from a built-in name.

– Describes the trigger but not the warning's exact format v0.4.2

15 Python HTTP client for batch predictions NEW 32

Adds a standalone Python HTTP client, later extended with Python HTTP batch predictions.

– Mentioned only in release summaries, no endpoint or method named v0.4.13 v0.4.9

16 Opt-in detection reasons and source diagnostics NEW 30

Adds opt-in detection reasons and source diagnostics to the language layer.

– No flag name or mechanism given beyond the label v0.4.13

WAS THIS USEFUL?



AutoTrust-JEV-27B-v1.1

Feedback



released · Oct 20, 2026

MODEL ↗

AutoTrust-JEV-27B-v1.1 is a decision model from AutoTrust.

AutoTrust-JEV-27B-v1.1 from AutoTrust was released on 17 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Reflection Beam 1.0 501B A23B

Feedback



released · Oct 20, 2026

MODEL ↗

Beam 1.0 501B A23B is a reasoning model from Reflection.

Beam 1.0 501B A23B from Reflection was released on 16 Oct 2026. 1049K-token context; \$0.25 in / \$1.00 out per million tokens (OpenRouter).

WAS THIS USEFUL?



FINAL-Bench Darwin-27B-ZTC-v9

Feedback



released · Oct 20, 2026

MODEL ↗

Darwin-27B-ZTC-v9 is a decision model from FINAL-Bench.

Darwin-27B-ZTC-v9 from FINAL-Bench was released on 17 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



FINAL-Bench Darwin-35B-ZTC-Vision

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

Darwin-35B-ZTC-Vision is a decision model from FINAL-Bench.

Darwin-35B-ZTC-Vision from FINAL-Bench was released on 17 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Liquid AI LFM2.5-24B-A2B-DSpark

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

LFM2.5-24B-A2B-DSpark is a chat model from Liquid AI.

LFM2.5-24B-A2B-DSpark from Liquid AI was released on 19 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



FrontiersMind Lumma-fev-27B

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

Lumma-fev-27B is a decision model from FrontiersMind.

Lumma-fev-27B from FrontiersMind was released on 13 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Moondream 3 Preview

Feedback



released · Oct 20, 2026

MODEL ↗

Moondream 3 Preview is a multimodal model from Moondream.

Moondream 3 Preview from Moondream was released on 15 Oct 2026. 32K-token context; \$0.30 in / \$2.50 out per million tokens (OpenRouter).

WAS THIS USEFUL?



Moondream 3.1

Feedback



released · Oct 20, 2026

MODEL ↗

Moondream 3.1 is a multimodal model from Moondream.

Moondream 3.1 from Moondream was released on 15 Oct 2026. 32K-token context; \$0.30 in / \$2.50 out per million tokens (OpenRouter).

WAS THIS USEFUL?



Meta Muse Spark Open

Feedback



released · Oct 20, 2026

MODEL ↗

BLOG ↗

Muse Spark Open is a reasoning model from Meta.

Meta: "Muse Spark Open is the most capable US open-weight model to date, and brings leading agentic capabilities to your desktop on Apple M5 Ultra Mac Studios."

Muse Spark Open from Meta was released on 15 Oct 2026. 1049K-token context; \$1.25 in / \$4.25 out per million tokens (OpenRouter).

WAS THIS USEFUL?



OmniJev OneJev1.5-4B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

OneJev1.5-4B is a decision model from OmniJev.

OneJev1.5-4B from OmniJev was released on 13 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



OmniJev OneJev1.6-0.8B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

OneJev1.6-0.8B is a decision model from OmniJev.

OneJev1.6-0.8B from OmniJev was released on 16 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



OmniJev OneJev1.6-27B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

OneJev1.6-27B is a decision model from OmniJev.

OneJev1.6-27B from OmniJev was released on 16 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



OmniJev OneJev1.6-4B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

OneJev1.6-4B is a decision model from OmniJev.

OneJev1.6-4B from OmniJev was released on 16 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



OmniJev OneJev1.6-9B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

OneJev1.6-9B is a decision model from OmniJev.

OneJev1.6-9B from OmniJev was released on 16 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Mistral AI Timestral-123M

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

Timestral-123M is a time-series forecasting model from Mistral AI.

Timestral-123M from Mistral AI was released on 14 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Mistral AI Timestral-209M

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

Timestral-209M is a time-series forecasting model from Mistral AI.

Timestral-209M from Mistral AI was released on 14 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Mistral AI Timestral-28M

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

Timestral-28M is a time-series forecasting model from Mistral AI.

Timestral-28M from Mistral AI was released on 14 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Mistral AI Timestral-5M

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

Timestral-5M is a time-series forecasting model from Mistral AI.

Timestral-5M from Mistral AI was released on 14 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



cribl-decision-1.0-27B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

cribl-decision-1.0-27B is a decision model from Cribl.

cribl-decision-1.0-27B from Cribl was released on 19 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



cribl-decision-1.0-9B

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

cribl-decision-1.0-9B is a decision model from Cribl.

cribl-decision-1.0-9B from Cribl was released on 18 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



vLLM Project d3-turbo

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

d3-turbo is a decision model from vLLM Project.

d3-turbo from vLLM Project was released on 19 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Mapika decider-12b-vision

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

decider-12b-vision is a decision model from Mapika.

decider-12b-vision from Mapika was released on 20 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



IBM granite-timeseries-ensemble-r1

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

granite-timeseries-ensemble-r1 is a time-series forecasting model from IBM.

granite-timeseries-ensemble-r1 from IBM was released on 23 Sep 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Perplexity API pplx-decider-flash-v1.2

[Feedback](#)[released · Oct 20, 2026](#)[MODEL ↗](#)

pplx-decider-flash-v1.2 is a decision model from Perplexity API.

pplx-decider-flash-v1.2 from Perplexity API was released on 15 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Perplexity API pplx-decider-v1.2

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

pplx-decider-v1.2 is a decision model from Perplexity API.

pplx-decider-v1.2 from Perplexity API was released on 15 Oct 2026. [Open weights](#) on Hugging Face.

WAS THIS USEFUL?



Shisa.AI shisa-de-2.1

[Feedback](#)

released · Oct 20, 2026

MODEL ↗

shisa-de-2.1 is a decision model from Shisa.AI.

shisa-de-2.1 from Shisa.AI was released on 19 Oct 2026. [Open weights](#) on Hugging Face.

Succeeds shisa-de-2.

WAS THIS USEFUL?



manjunathshiva OpenDecider

Feedback



v0.9.0

NOTES

RANK

Open, calibrated System 1 decision models: typed choice / score / yes-no answers with a probability for every option. 400M to 80B; CPU, NVIDIA, Apple Silicon, LM Studio, Ollama and vLLM.

OpenDecider v0.9.0 adds Microsoft-Decision-1 benchmark results across its three supported benchmarks.

WHAT SHIPPED • 1 FEATURE most completely described first ? what's the number?

01 Microsoft-Decision-1 benchmark results added NEW 20

OpenDecider v0.9.0 adds evaluation results for Microsoft-Decision-1 across all three supported benchmarks.

– Names the model and benchmark count but no mechanism or scores

v0.9.0

WAS THIS USEFUL?



PrismML builds 1-bit and ternary Bonsai models that pack advanced reasoning, tool use, and multimodal AI into phones, laptops, and efficient data centers.

PrismML expanded its Bonsai 2 family with new 2B, 4B, and 9B parameter on-device models combining reasoning, vision, and tool calling.

↳ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 Bonsai 2 family expanded with 2B, 4B, 9B models NEW

38

PrismML added Bonsai 2 models in 2B, 4B, and 9B sizes, each supporting reasoning, vision, and tool calling for on-device AI agents.

– Names model sizes and capabilities but no mechanism or usage detail

PrismML Expands Bonsai 2 Family with 2B...

WAS THIS USEFUL? 👍 👎

StartLux

Feedback



v1.1.0

NOTES

RANK

Open decision models from 0.8B to 27B: a calibrated probability for every option in one pass, and a short thought only when unsure. Engine for vLLM, llama.cpp and MLX.

StartLux-Decision v1.1 introduces a unified multi-backend inference engine for its System One choice-scoring format, adds per-choice uncertainty reporting and a tunable confidence gate, and expands its model lineup with new AutoThink reasoning models alongside Instant models across sizes.

WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 Unified `startlux-decide` inference engine NEW 73

A new engine, `startlux-decide`, supports vLLM, llama.cpp, and MLX backends and serves the System One `/v1/systemone` request format, giving a single interface across these three runtimes.

– Names backends and endpoint, no usage example given. v1.1.0

02 AutoThink and Instant model lineup NEW 70

New AutoThink models at 4B, 9B, and 27B answer in one pass when confident or prepend a short reasoning step when not; alongside Instant models at 0.8B, 2B, 4B, 9B, and 27B, both are published on Hugging Face with GGUF and MLX builds that support image input.

– Sizes and formats listed, no usage example or benchmark. v1.1.0

THINNER COVERAGE BELOW

03 Per-request confidence gate threshold NEW 58

Each request can set its own confidence gate threshold, defaulting to `0.6`, via per-request configuration, controlling when the model falls back to reasoning.

– Default value given but no example request payload. v1.1.0

04 `none\_probability` in choice responses NEW 50

The `/v1/systemone` response now reports a `none_probability` value alongside every choice answer, surfacing model uncertainty per option.

– Field and endpoint named but no sample response shown. v1.1.0

WAS THIS USEFUL?



Shisa.AI Shisa

Feedback



Shisa DE-2: An Open Decision Model

NOTES

RANK

Shisa.AI released Shisa DE-2, an open-source decision model that returns calibrated probabilities for yes/no and multiple-choice questions instead of free text.

Shisa.AI released Shisa DE-2, an open-source decision model that returns calibrated probabilities for yes/no and multiple-choice questions instead of free text.

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

01 Shisa DE-2 decision model NEW

64

Shisa DE-2 is an Apache 2.0 licensed decision model built on Gemma 4 26B-A4B that answers yes/no and multiple-choice questions by returning probabilities rather than generated text. It scores 59.01 on the public Decision Index 0.3 benchmark with a median response time of 27 ms.

– Names model, license, benchmark score and latency; no API or usage command given.

Shisa DE-2: An Open Decision Model

WAS THIS USEFUL?



// END OF TRANSMISSION

The AI Toolchain · curated daily · October 22, 2026

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

Subscribe

Feedback on this issue