

THE TAKE

The useful overnight changes are in the plumbing. Local inference gains more GPU paths, code graphs lose misleading edges, and video export gets finer controls. Each comes with boundaries: hardware fit, prerelease status, or compatibility changes still decide what is ready to use.

Coverage: 31 October 17:30 to 1 November 2026, 05:30 UTC. Earlier missed items and scheduled changes are explicitly dated.

[Agent tools / 2](#) · [Costs and operations / 3](#) · [Creator tools and security / 4](#)

LOCAL AI / RELEASED 1 NOV, 00:39 UTC

Runner moves more local models onto the GPU

- Runner 1.10.0 adds whole-model Metal execution for dense Qwen3.5-family models, dense Granite-4 h and Nemotron-H. The entire model must fit: partial offload, adapters and several mixture-of-experts paths still fall back to CPU.
- On CUDA, NVFP4 tensor-core execution becomes the default for prompt processing. This does not change decoding, and Ampere uses a dequantizing route rather than native FP4 tensor cores. The release is not a universal speed boost.
- Its new quantization-error tool and sensitivity planner measure tensor error per byte. The planner has been compared with float32 on only one model so far, so treat it as a heuristic rather than a general quality guarantee.
- The notes also correct earlier prefix-cache claims: reused prompts are bit-identical to a cold run only with matching chunk geometry. Near-ties can produce a DIVERGED receipt. Open-source downloads are available; model licences and memory requirements remain separate.

[Runner 1.10.0 release and limitations](#) · [Project overview and downloads](#)

PRERELEASE BUILDS / 31 OCT, 18:42-21:35 UTC

llama.cpp fixes image preprocessing and agent endpoints

- Builds b12021-b12024 remove letterbox padding from Gemma4 image resizing to match Transformers preprocessing, honor extra_params such as seed and min_p on /v1/responses, and repair a multi-slot /v1/systemone HTTP 500 failure.
- An LFM short-convolution change targets batched decoding. The maintainer reports mostly batch-size-dependent gains on H100 and MI325X; this is not evidence of a general speedup on a personal computer. All four builds are prereleases, distinct from yesterday's OpenCL update.

[b12021 / image input](#) · [b12022 / Responses](#) · [b12023 / multi-slot fix](#) · [b12024 / LFM](#) · [Batched-decode method and measurements](#)

AGENT TOOLS / NEW FIXES AND A DATED CATCH-UP

MISSED CATCH-UP / RELEASED 30 OCT, 23:27 UTC

Pydantic AI tightens fallback and state handling

- Version 2.64.0 binds stored response, item, conversation and container IDs to the endpoint that created them. Integrations that switch providers should review this compatibility change instead of assuming server-side state can travel with the conversation.
- Retrying transports now raise `ModelHTTPError` when retries are exhausted so a fallback can proceed. A first-event timeout handles streams that never start; sticky fallback can keep the conversation on the model that successfully answered.

- `ArtifactScanner` adds opt-in scanning for Memory and Skills; installing the update alone does not enable it, and MCP results are outside its scope. Shared-usage subagents now obey parent limits. These controls do not guarantee security or spending outcomes.
- Check migrations: Coder tool schemas are now non-strict by default, with `Coder(strict=True)` to opt back in; `json-repair` is optional except for Coder and `RepairToolArguments`. This is an earlier release missed by our previous editions, not a new overnight launch.

[Pydantic AI 2.64.0 release notes](#)

DEVELOPER TOOLS / 1 NOV, 00:53 UTC

Graphify fixes graphs and caches prompts

- Graphify 0.9.100 stops differently cased type and function names from colliding, a bug that could drop a node and invent a self-loop. Exported graphs also retain parallel relationships, and failed backups no longer count as successful.
- The Claude backend now caches its stable system prompt. Gateways that reject `cache_control` can disable this with `GRAPHIFY_CLAUDE_PROMPT_CACHE=0`. The notes promise no universal cost reduction; actual savings depend on provider caching and use.

[Graphify 0.9.100 release notes](#)

MISSED CATCH-UP / 31 OCT, 10:56 UTC

Excel agents can save the opening view

- `ExcelMcp` 2.6.6 adds window set-view for scroll position and optional selection before saving. Its readback reports the top-left cell, selection and active cell, helping generated dashboards open where intended.
- Available through CLI and MCP. It requires Windows, desktop Excel 2016 or later and a signed-in interactive desktop session; this is not a browser-only Excel service.

[ExcelMcp 2.6.6 release](#)

COSTS AND OPERATIONS

SCHEDULED CHANGE / 1 NOV / ANNOUNCED 1 OCT

Cloudflare AI Search reaches its billing date

- Cloudflare says usage-based AI Search billing begins November 1. Every account includes 5 million ingestion tokens, 10 GB-month of storage, 1,000 semantic queries and 1,000 full-text queries each month.
- Beyond those amounts: base ingestion is \$0.75 per million tokens, storage \$2 per GB-month, semantic/vector/hybrid queries \$0.75 per 1,000, and full-text queries \$0.10 per 1,000. Image processing, including OCR, adds \$0.50 per million ingestion tokens.
- Managed embeddings and reranking are included. Generation, query rewriting and external providers remain separately billed through their applicable services. Chunk overlap counts repeatedly toward ingestion, so document size alone does not predict usage.
- The pricing page gives a date, not an exact switch-on hour or timezone. We verified the published schedule, not a charged transaction; this is not a new general-availability announcement.

[Official pricing and included usage](#) · [1 October announcement](#)

UPDATE / 31 OCT, 22:32 AND 23:31 UTC

Northflank adds two more rotation tools

- Minio addon secret rotation is now available to all customers, followed by RabbitMQ. These are additions to the MySQL/Redis tools reported in the evening edition.
- The incident remains Identified / Major outage. The provider still says to rotate all secrets, revoke old credentials at their issuers, update workloads and inspect AI-provider usage. New rotation tools do not mean the incident is resolved.

[Northflank incident timeline](#)

CREATOR TOOLS AND SECURITY

MERGED 1 NOV, 03:00 UTC / NOT A STABLE RELEASE

ComfyUI adds ProRes export on its development branch

- ComfyUI merged ProRes/MOV video export and selectable chroma subsampling. The Create Video node exposes 4:4:4, 4:2:2 and 4:2:0 choices, giving editors more control over export formats rather than changing the generation model.
- Unsupported combinations fall back with a warning. The latest stable release remains October 27's v0.43.0; this is master-branch code, not a confirmed Desktop or Cloud rollout. We inspected the change, not a rendered export.

[Merged PR and implementation](#) · [Latest stable release](#)

MISSED CATCH-UP / DISCLOSED 30 OCT / COVERED 31 OCT

A breach disclosure, not proof of an autonomous attack

- Chainalysis confirms access to customer names, emails, user IDs and organization-role labels. Its investigation has found no evidence to date of access to investigation content, plaintext passwords, MFA secrets or session tokens; the scope remains under review.
- The firm attributes the intrusion to offensive agentic techniques, but has not explained the exploit path or how much was automated. Independent reporting confirms the disclosure, not the AI attribution. Customers should scrutinize unexpected authentication requests and follow the company notice.

[Chainalysis original disclosure](#) · [Independent reporting](#) / [BleepingComputer](#)

Edition notes

Nine deduplicated developments: overnight releases and code changes, one incident update, a scheduled billing date and clearly labelled catch-ups. No new frontier-model launch or significant maker-hardware release was verified. The opening take is editorial interpretation.

PUBLIC COVERAGE

Research covered public/indexed X, original release notes, repository changes and credible news. Direct X pages were not reliably accessible; no inaccessible post is treated as proof. This is neither a personal feed nor exhaustive X coverage.

EVIDENCE LIMITS

Release capabilities are maintainer claims, not our performance tests. Exact times are UTC where verified; day-only changes retain that precision. Compatibility and access limits matter more here than headline performance claims.