The Tech Times

MONDAY 2 NOVEMBER 2026 MORNING / 05:30 GMT

THE TAKE

Better models still need careful plumbing. LongCat brings a new paid multimodal API, while fresh infrastructure fixes and dated catch-ups reveal the operational work around it: compute selection, agent permissions, inference paths and security previews whose limits matter as much as their headline capabilities.

Coverage: 1 November 17:30–2 November 05:30 UTC, 2026. Three new-window items; seven dated catch-ups absent from the prior archive. Opening take: editorial interpretation.

Local models and MCP / 2 · Creator tools / 3 · Developer operations / 4

MODELS / API NOTICE 2 NOV / REPORT 01:18 UTC

LongCat-2.5 moves from preview to a paid multimodal API

- Meituan's final LongCat-2.5 combines visual understanding with long-running coding and agent tasks. The maker describes a 1.6-trillion-parameter mixture-of-experts model with about 48 billion parameters active per token. This follows September's Preview; it is not that older release relabelled.
- The current API lists a 1,048,576-token context and up to 262,144 output tokens. Official guides cover Claude Code, Codex, OpenCode and Hermes. Access requires an account, key and credit balance; the announcement says downloadable weights are coming, not already available.
- Limited-time API rates per million tokens are $0.30 uncached input, $0.006 cached input and $1.20 output. No discount end date is given. Published benchmark gains are vendor results, not independent proof that it will outperform your current model.

Official availability notice · Current pricing · Context and output limits · Dated independent report

BETA / RELEASE ENTRY 2 NOV / STAGED ROLLOUT

MLflow can select compute for viewing traces

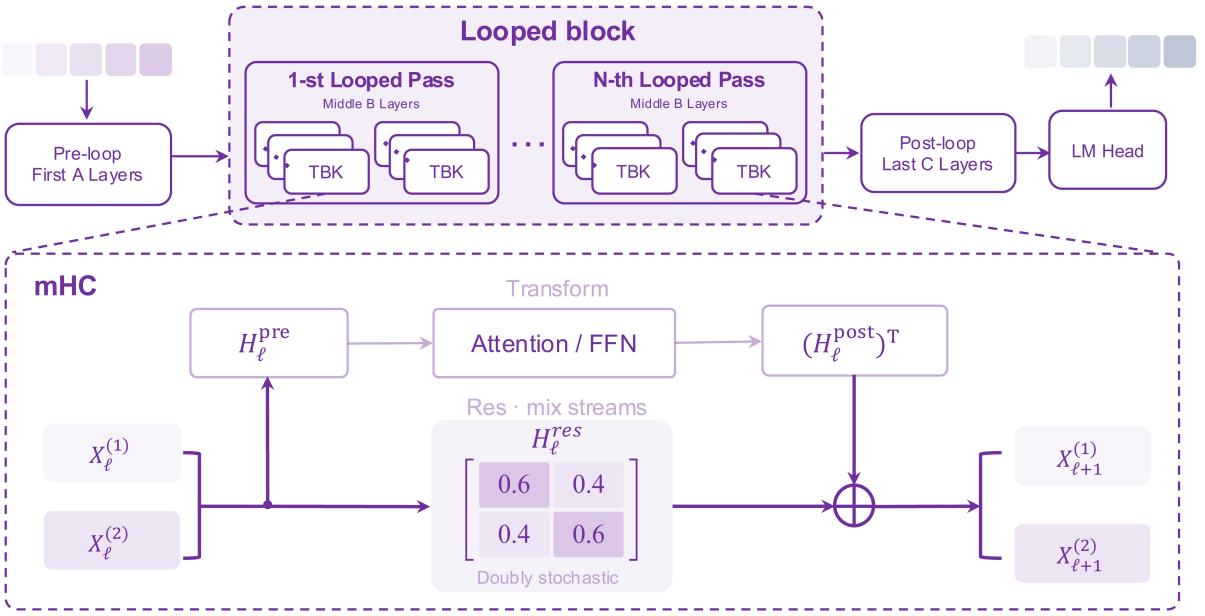
- Databricks lists a managed-warehouse beta that removes manual SQL-warehouse selection when browsing Unity Catalog traces in the MLflow UI. An existing selected warehouse takes priority; otherwise the path depends on Lakehouse//RT beta and serverless availability.
- An administrator must opt in, users need SELECT access to the trace tables, and the region must support serverless. Normal compute charges apply. Python SDK access still requires a SQL warehouse. The release is staged, so the date is not proof it is available in every account.

November release notes · Beta requirements and billing

LOCAL MODELS AND MCP / DATED CATCH-UPS

OPEN-WEIGHT CATCH-UP / PREVIEW / 30 OCTOBER

SenseNova reuses layers for a deeper execution path



The same middle block runs more than once. Original architecture diagram; it explains weight reuse, not measured speed or memory savings. SenseNova / official model card

- SenseNova-Loop-9B-Preview accepts text and images and generates text. Its architecture runs the same 24-layer block twice: 32 distinct decoder layers produce 56 layer executions. The point is a longer computation path without storing a second copy of those weights.
- Weights are available under Apache 2.0, but deployment is not a generic drop-in: the model card requires matching custom execution code and points SGLang/vLLM users to specific pull-request implementations. Its training sequences reach 262,144 tokens; this is not a guarantee of fast consumer-GPU inference.

Dated model card, weights and deployment limits

MCP CATCH-UP / 1 NOVEMBER / TIME UNSTATED

GetMCP groups selected servers behind one sign-in

- GetMCP 1.8 adds project gateways for up to 50 servers, per-tool switches for remote MCP servers, and imports of client configurations such as Codex TOML. Conversion replaces secrets with placeholders. Per-client keys can expire or be revoked.
- The release also says gateways no longer forward a caller's credential to member servers, with stronger OAuth binding and upstream address checks. These are vendor-described safeguards, not a security audit. It runs as a self-hosted PHP app or WordPress plugin; feature-specific plan limits were not verified.

Release and security changes

CREATOR TOOLS / UPDATE AND DATED CATCH-UPS

CREATOR UPDATE / MASTER MERGE 2 NOV, 04:02 UTC

ComfyUI fixes an FP8 path that silently ran in BF16

- ComfyUI merged a fix for fp8_e4m3fn_fast when dynamic VRAM is enabled. A BF16 checkpoint could silently use BF16 computation instead of the intended FP8 matrix-multiply path; the patch checks the dtype used when weights reach the GPU.
- The contributor reports 1.08 to 1.73-1.83 iterations/second on one RTX 5070 Ti 16GB FLUX.1-dev test. That is a contributor result, not our benchmark or a universal speed-up. Stable remains v0.43.0 from October 27; this fix is currently on master.

Merged PR and contributor test · Implementation · Current stable

---

CREATOR CATCH-UP / RELEASED 26 OCT, 00:25 UTC

Waytator puts recording and quick edits in one Wayland app

- Waytator 1.5.0 adds region, screen and window recording with system sound and microphone input, then lets you trim, crop, split, join and adjust the clip in the same app. Exports include H.264, HEVC, AV1, VP9 and GIF - a practical local tool for short demonstrations and bug reports.
- This is an October 26 release, not an overnight launch. The GPL-3.0-or-later app targets supported Wayland/wlroots desktops, with Arch AUR, Nix and source-build routes. The current README warns that copied videos are temporary: use Export to retain them. Nixpkgs can lag releases; desktop compatibility and output quality need testing.

Official release · Installation, screenshots and limitations

POLICY CATCH-UP / STAFF POST 30 OCT, 16:58 UTC

Claude Code draws a line at specific game recreations

- Anthropic’s Addy Osmani confirmed a prompt change steering Claude Code away from recreating someone else’s specific existing game. Original games, genre-based games and work using the user’s own designs and assets are intended to remain supported. This is not a ban on game development, nor legal clearance for any particular project.

Anthropic staff explanation on X

DEVELOPER OPERATIONS / DATED CATCH-UPS

SECURITY CATCH-UP / ANNOUNCED 28 OCTOBER

GitHub scans agent data paths, with narrow preview limits

- GitHub's AI-app scanning preview looks for risky paths from untrusted content into agents. The announcement names several AI risks, but the current coverage section limits the preview to indirect prompt injection. This is a separate detector from CodeQL and conventional AI Scan.
- It is off by default, requires GitHub Advanced Security and consumes additional AI credits, including for eligible public repositories. Findings are advisory: they do not block merges. It scans pull requests, not the whole backlog, and excludes fork and Dependabot PRs.

Announcement · Current coverage, billing and limits

CODING CATCH-UP / 1 NOVEMBER, 03:22 UTC

Qwen Code adds bounded history and a retry budget

- Qwen Code 0.26.5 enables Generated UI by default: interactive card values can be attached as removable context to the next message. Bounded SQL history queries and paged older messages aim to keep long managed sessions manageable.
- The release also caps autofix-pr at three fix pushes before handing a failing check to a human. Its in-place PR-watch option can push branches and reply on pull requests; that is a consequential workflow choice, not just a viewing mode. A one-command managed-agent Helm deployment is explicitly experimental.

Release notes and feature limits

DEVELOPER CATCH-UP / ANNOUNCED 28 OCTOBER

One model ID can now change behind your app

- Vercel's virtual models move the selected model, fallback order and settings behind a reusable vmc/ identifier. Teams can change that configuration centrally without redeploying every app or reconfiguring every agent.
- Central settings override conflicting request settings. Provider-specific options do not carry across automatically when switching vendors; run evaluations against a candidate configuration first. Vercel says Gateway passes through provider pricing without an inference markup or platform fee.

Launch and operational caveats · Dated changelog

PUBLIC COVERAGE

Research uses public/indexed X, official announcements, repositories, documentation and credible reporting—not your personal feed or exhaustive X coverage. The dated Anthropic staff post was verified via its public syndication record.

EVIDENCE AND DATES

Catch-ups retain their actual dates, not an aggregator's current date. Capabilities are source claims, not our hands-on tests. The ComfyUI speed figure is one contributor test; previews and staged rollouts are explicitly labelled.