

MORNING EDITION — THURSDAY 15 OCTOBER 2026

Reporting window: 14 Oct 2026, 17:59 CEST to 15 Oct 2026, 06:10 CEST (Paris time). Issued 15 Oct 2026, 06:10 CEST.

Two overnight software releases clear the verification cutoff after the previous actual edition, issued October 14 at 17:59 CEST: Claude Code 2.1.300 and Ollama 0.40.5-rc0. We lead with their concrete changes rather than repeat yesterday's model announcements. A separate October 9 physics preprint is labeled BACKGROUND, with its limitations and the community's contrasting reactions. Broader model and policy leads that could not be dated or verified are omitted.

2 NEW OR UPDATED DISPATCHES · 1 BACKGROUND ITEM · PREVIOUS: [2026-10-14-AFTERNOON](#)

THE NEWS

CODING & AGENTS NEW

CLAUDE CODE 2.1.300 FIXES LOST RESUMED CONTEXT AND TIGHTENS HOOK VALIDATION

PUBLISHED: 14 Oct 2026, 20:24 CEST

Anthropic published Claude Code v2.1.300 at 20:24 CEST on October 14, after the previous edition. Its release notes describe fixes for headless and Agent SDK sessions losing recent messages or compaction summaries when resuming after interrupted, retried or output-limited replies.

The release also validates updatedInput returned by PreToolUse and PermissionRequest hooks: unknown keys are dropped, and input rejected by the tool denies the call. Local session transcripts now also record each API request's system prompt, tools and parameters, increasing the information present in a transcript a user chooses to share.

Other additions include separate five-minute and one-hour cache-write token counts in OpenTelemetry, GenAI trace attributes, and the /advisor feature on Claude Platform on AWS. An advisor does not run until selected.

ANALYSIS — WHY IT MATTERS. *For unattended agents, preserving the right context after a restart and rejecting malformed hook inputs are operational safeguards, not just interface polish. Review transcript contents before sharing them.*

CAVEAT. These are the maintainer's release claims, not independent regression-test results. This issue reports the software release; it does not claim a new Claude model or install the update.

SOURCES

- Anthropic: Claude Code v2.1.300**
<<https://github.com/anthropics/claude-code/releases/tag/v2.1.300>> (primary source) **CHECKED: 15 Oct 2026, 06:08 CEST**
- GitHub API: original release timestamp**
<https://api.github.com/repos/anthropics/claude-code/releases?per_page=5> (primary source) **CHECKED: 15 Oct 2026, 06:05 CEST**

INDUSTRY & POLICY NEW

OLLAMA RELEASE CANDIDATE PORTS MLX CUSTOM KERNELS TO CUDA

PUBLISHED: 14 Oct 2026, 21:45 CEST

Ollama published v0.40.5-rc0 at 21:45 CEST on October 14. The release candidate includes a port of MLX custom kernels from Metal to CUDA and support for custom tool calls in OpenAI-compatible Responses history.

The linked implementation covers mamba2_scan, depthwise_conv_silu and gated_delta kernels that previously fell back to graph operations on CUDA. Its author reports testing on a Linux RTX 6000 Ada, with Nemotron prompt-processing improvements of 110% and 122% for two tested NVFP4 configurations and generation improvements of 33% and 15%, respectively.

The code change was merged on October 13; the fresh development is its inclusion in this October 14 release candidate, not a newly invented kernel announcement. The preceding stable v0.40.4 release predates the current news window.

ANALYSIS — WHY IT MATTERS. *More efficient CUDA paths can make local inference more practical on NVIDIA hardware, while Responses-history compatibility matters for tool-using agents. The reported speedups are useful leads for testing, not universal performance guarantees.*

CAVEAT. This is explicitly a prerelease. Performance figures are author-reported for two model configurations on one GPU; they have not been independently reproduced here and do not apply to every model or device.

SOURCES

- Ollama: v0.40.5-rc0 prerelease**
<<https://github.com/ollama/ollama/releases/tag/v0.40.5-rc0>> (primary source) **CHECKED: 15 Oct 2026, 06:08 CEST**
- GitHub API: original release timestamp**
<https://api.github.com/repos/ollama/ollama/releases?per_page=3> (primary source) **CHECKED: 15 Oct 2026, 06:06 CEST**
- Ollama #18657: implementation and author benchmark**
<<https://api.github.com/repos/ollama/ollama/pulls/18657>> (primary source) **CHECKED: 15 Oct 2026, 06:08 CEST**

BACKGROUND — OLDER OR UNDATED CONTEXT

Earlier or undated items, carried for context only. They are not new reporting.

RESEARCH	BACKGROUND — OLDER CONTEXT
BACKGROUND: THEOPHYSEVAL REPORTS 98% ON 50 PHYSICS CALCULATIONS, NOT AUTONOMOUS DISCOVERY	
PUBLISHED: 9 Oct 2026	
BACKGROUND - October 9, 2026. The TheoPhysEval preprint assembled 50 unpublished, verifiable theoretical-physics problems. Its authors report 98% final-answer accuracy for GPT-6-Astra in Codex with web and computational tools, at an average cost below US\$3 per problem. This is an older paper newly examined by this journal, not an overnight model launch. The authors describe the problems as concrete calculations using known techniques, not breakthrough open questions. GPT-6-Astra and Opus 5.5 in Claude Code each scored 92% on the isolated track; DeepSeek-V4.1-Flash in Codex scored 54% with web access and 50% without it. These results describe the tested model-and-harness combinations, not a ranking for all work. The October 14 discussion mixed excitement with caution. u/141_1337 wrote, "Next year is gonna be physics, I can feel it lol"; u/quivering_palm answered the broader optimism with, "Not really, physics will be bottlenecked by the need to run IRL experiments, which take a lot of money and time." These are anecdotal reactions, not evidence of a new scientific discovery.	
ANALYSIS — WHY IT MATTERS. <i>The result supports evaluating agents on well-specified research calculations. It does not show that they can choose important scientific questions, design experiments or autonomously produce major new theories.</i>	
CAVEAT. This is a preprint, with final-answer-based grading and only 50 problems. Some human contributors used AI when solving problems, so the authors cannot rule out data leakage. They also discuss literature memorization; the benchmark is not an adversarial test of original scientific invention.	
COMMUNITY REACTION (OPINION, NOT VERIFIED FACT) “Next year is gonna be physics, I can feel it lol” — u/141_1337 (244 upvotes when checked) “Not really, physics will be bottlenecked by the need to run IRL experiments, which take a lot of money and time.” — u/quivering_palm (52 upvotes when checked)	
SOURCES 1. <u>arXiv: TheoPhysEval abstract and October 9 submission</u> <https://arxiv.org/abs/2610.16733> (primary source) CHECKED: 15 Oct 2026, 06:05 CEST 2. <u>TheoPhysEval: methods, results and limitations</u> <https://arxiv.org/html/2610.16733v1> (primary source) CHECKED: 15 Oct 2026, 06:06 CEST	

KEY PATTERNS — EDITORIAL ANALYSIS

— The fresh releases improve agent-state reliability, hook validation and inference plumbing rather than establish a newly launched foundation model. — A CUDA speedup from one author’s two-model test is narrower evidence than a general performance claim; Ollama’s release remains a candidate. — High accuracy on specified physics calculations is not the same as autonomous discovery; the October 9 paper itself makes that distinction.

SOURCE COVERAGE & COLOPHON

COLLECTION ENGINE	SOURCE STATUS
VERSION: 3.28.0 RAN THIS EDITION: yes ENGINE LOOK-BACK: 2 days	REDDIT: ok YOUTUBE: ok HACKERNEWS: ok POLYMARKET: unreachable GITHUB: ok

DEVTO: ok
x: skipped-unconfigured

COVERAGE NOTES

- Two verified software releases appeared after the previous actual issue at 17:59 CEST on October 14. One October 9 physics paper is explicitly BACKGROUND; it is not counted as fresh news.
- Models, coding and agents, research, and industry, infrastructure and policy were searched over a two-day discovery window. No separate new model launch or policy change met the verification gate; this does not establish a quiet wider news cycle.
- The research engine completed with 69 evidence items: 16 Reddit threads, 6 YouTube videos, 20 Hacker News stories, 14 GitHub items and 13 dev.to articles. These are evidence items, not 69 verified developments.
- Polymarket was unreachable because TLS hostname verification failed. No odds are reported and TLS verification was not bypassed. X was not configured.
- YouTube returned six videos but transcripts for only two; the other transcript lanes were rate-limited or unavailable. The overall engine source state is preserved as ok alongside this limitation.
- Three post-engine web supplements were attempted. The configured search backend failed and its built-in keyless fallback returned leads for one query and empty results for two; empty searches are not evidence of no news.
- Reuters ASML and Axios policy article verification remained blocked. Date-only or older reports, commentary, unsupported social claims and a content-free Codex alpha release were not promoted as fresh stories.
- No browser cookies, new optional services, credentials, paid enrichment, TLS bypass or schedule changes were used. Raw research and provenance remain private; only sanitized newspaper files are eligible for the authorized public repository.

THIS EDITION

- [Exact input JSON](#)
- [Markdown text](#)
- [Permanent page](#)

ENGINE FOOTER

All agents reported back!
└─ Reddit: 16 threads | 2,016 upvotes | 605 comments
└─ YouTube: 6 videos | 676,114 views | 2/6 with transcripts
└─ HN: 20 stories | 2,140 points | 1,550 comments
└─ dev.to: 13 articles | 286 reactions | 152 comments
└─ GitHub: 14 items | 41 reactions | 194 comments
└─ Top voices: r/singularity, r/artificial, r/LocalLLaMA

The AI Gazette is a locally rendered digest. Each dispatch separates reported facts, labeled analysis and caveats, and cites the sources checked at the times shown. Background items are older or undated context, not new reporting.

Rendered as static HTML with inline styles and no scripts, cookies or analytics. Typefaces are requested from Google Fonts (fonts.googleapis.com, fonts.gstatic.com); if that request is blocked, Georgia, Times or another installed serif is used.