

THE MORNING TECH BRIEF

Sunday, October 4, 2026 · Edition No. 5

Llama-server learns to decide, Strix Halo eats 300B MoEs, and RuneScape finally gets a number four.

AI & AGENTS

[llama.cpp lands /v1/systemone in llama-server — typed decision models go local](#)

Your weekend calibration project just landed in the ecosystem itself: Jev-style decision models (Laya, Julia-1, Kev, Bespoke Nimble added in b11364) now serve locally with typed answers. Ollama's decision API has competition — pick whichever is faster on the A770.

[Aleph Alpha open-sources Kolibri — 78.1B MoE, Apache 2.0](#)

Thinks like a 3.5B, eats 78GB of memory like a 78B — the MoE tradeoff at its most honest. FP8 rules it out for your 16GB A770, but the community already has it hobbling on 128GB CPU boxes. File under “what could be.”

[llama.cpp merges MTP speculative decoding for Qwen3.8-Flash-Next](#)

Free speed on your local Qwen lane: the checkpoint's own multi-token prediction head now drives speculative decode, with a claimed ~55% throughput bump. Treat it as “try it on your box” rather than gospel.

[PewDiePie ships Ajax — a fine-tuned Qwen 3.5 9B local agent](#)

A 9B Qwen doing web, email, and calendar locally ~90% of the time is your Hermes thesis in the wild — plus open-source Heretic ablation tooling and OpenAI banning him twice for distillation. The drama-to-signal ratio is immaculate.

HARDWARE

[Two ~300B MoE models run on a single 128GB Strix Halo mini PC](#)

GLM-5.3-Flash at ~580 tok/s prefill, MiMo-V2.6-Flash at 44 tok/s decode via EXL3 + the open Kyojin ROCm engine — all on a desk box. Unified memory keeps resetting the “how big a model fits at home” bar, and AMD keeps holding the eraser.

[Minisforum MS-S1 Max hits UK shelves — Strix Halo with a real PCIe x16 slot](#)

The one 128GB Strix Halo box that isn't a closed loop: a physical x16 slot means bolt a discrete card in later for hybrid inference. 64GB starts at £2,359 — your N95 nomad box's rich cousin.

[Intel raises PC CPU prices 10% tomorrow, Oct 5 — third hike this year](#)

CEO Lip-Bu Tan says memory prices are up 5–7x and Intel can't fill half its CPU orders — datacenters are eating your build budget. If you're buying Intel silicon or DDR5/NAND, the clock is literally ticking to tomorrow morning.

[NVIDIA quietly raised the 128GB DGX Spark to \\$6,950 — 74% in a year](#)

The 128GB Spark went from \$3,999 to \$6,950 while a 128GB Strix Halo box does the same model-holding job at \$2,959–\$3,649 without CUDA. If you don't need the NVIDIA software stack, the math now points the other way, hard.

SOFTWARE

[Python 3.15.0rc3 ships — final lands October 9](#)

A surprise third RC because lazy imports (PEP 810) tripped blockers — but the ABI freeze is the news: wheels you build now ship clean, and the beefier JIT + Tachyon sampling profiler make this the meatiest Python in years.

[Wine 11.19: Wayland color management for Vulkan, DNS caching, Unicode 18](#)

The headliner is Wine's Wayland driver implementing color management for Vulkan — color-accurate Windows games on Wayland, the "HDR without XWayland color weirdness" fix your Ark setup would feel. 23 bug fixes and 350 commits besides.

Zen Browser 1.23b: Library hub, Firefox 157 base, 70+ vuln patches

New Library pane and hardware AV1 decode for WebRTC — but the mandatory item is 70+ high-impact fixes from MFSA2026-97. Upgrade-now release, not a feature sampler.

Kagi kills Orion for Linux and Windows

The last non-Chromium indie browser bet folds on desktop — macOS/iOS only, source promised within 30 days, no maintainer lined up. The WebKit-on-Linux dream dies in a repo nobody ships.

GAMES

Jagex announces RuneScape 4 — first numbered sequel in 13 years, built in Unreal

Grown out of Dragonwilds' expansion plans, RS3/OSRS aren't going anywhere, and it's "a few years away" — but the biggest MMO news of the weekend, no contest.

Nine games leaving Xbox Game Pass mid-October — including 2025's GOTY

Clair Obscur: Expedition 33 gets evicted around Oct 15 (discounted if you buy), plus Pacific Drive, A Plague Tale: Requiem, and Crime Scene Cleaner. Finish Expedition 33 first — trust me.

s&box free to play on Steam this weekend (Oct 3–5)

Facepunch's Garry's-mod-on-steroids sandbox is free until Monday 10am PT, 20% off if it hooks you. Go build something cursed.

GITHUB WATCH

antirez/ds4 — DwarfStar: narrow, self-contained C inference engine

The Redis guy's thesis is "refuse to be general" — one engine, DeepSeek V4 Flash/PRO + GLM 5.x + Qwen 3.8 Flash Next, Metal/CUDA/ROCm, built-in HTTP server. The polar opposite of Ollama/llama.cpp — and he's winning on the numbers. ~23k stars.

thedotmack/claude-mem — persistent memory across agent sessions

Hooks the session lifecycle, compresses tool observations with AI, re-injects only relevant context — claims ~75% token savings. The compat list literally includes Hermes. Cross-session memory is the layer your local Hermes loop is missing. (~76.5k stars; AGPL 3.0, fine for the home lab.)

AlexsJones/llmfit — one command to see what fits your hardware

Auto-detects CPU/RAM/GPU/VRAM, scores hundreds of models on quality/speed/fit/context, picks optimal quant. TUI + CLI + REST. Settles the "will it fit in 16GB VRAM" debate in one command — and ships as a skill for agent harnesses.

VectifyAI/PageIndex — vectorless, reasoning-based RAG

Hierarchical tree index + LLM reasoning instead of chunk-and-embed; claims 98.7% on FinanceBench vs ~50% for vector RAG. No Pinecone, no FAISS, no chunking tax — a local-friendly doc-QA engine with MCP integration for Hermes.

earthtojake/text-to-cad — coding agents get CAD superpowers

Text-to-parametric-CAD code generation with geometry validation, pip installable. +1,727 stars today. PI-DECK is a 3D-printed project and you're hand-drawing engineering figures — an agent that generates validated CAD closes the loop from "Hermes decides" to "Ender 3 prints." The most Matt-shaped repo on trending right now.

Sources: MarkTechPost, rohital.com, groundtruth.day, AI Weekly, Techyon, Camerajabber, StartupFortune, ComputingForGeeks, Python blog, Phoronix, LinuxCompatible, nerds.xyz, ScreenRant, ThisGenGaming, GameRant, GitHub trending. Research cut: the 24 hours to Sunday, October 4, 2026 · Built for Matt's stack: Omarchy/Arch, local Qwen, Nous cloud, Hermes harness.