MONDAY, MARCH 23, 2026 · ISSUE #49

The Rizz News

Yesterday's Top Tech Stories — Curated by RizzBot

Flash-MoE: Running a 397B Parameter Model on a Laptop

▲ 373 · 118 comments · github.com/danveloper

TL;DR: Flash-MoE allows a 397B parameter AI model to run on a MacBook Pro, achieving 4.4+ tokens/second with production-quality output through a C/Metal inference engine.

Researchers have successfully run a 397-billion parameter Mixture-of-Experts (MoE) model, Qwen3.5-397B-A17B, on a MacBook Pro with 48GB RAM, achieving over 4.4 tokens/second. This impressive feat was accomplished using a pure C/Metal inference engine, bypassing traditional Python frameworks. Key to this performance are techniques like on-demand SSD expert streaming for the 209GB model weights and FMA-optimized Metal compute shaders, enabling production-quality output with full tool calling capabilities directly on consumer hardware.

WHAT THE COMMUNITY SAYS

The discussion centers on the trade-offs of running a heavily quantized (compressed) 397-billion parameter AI model on powerful consumer hardware. One key perspective is that specific ~2.5 bits-per-weight (BPW) quants are highly capable, achieving impressive benchmark scores and enabling offline use. However, a countervailing view argues that these low-bit models are deceptive, failing in practical, long-session tasks and producing errors, making them a "waste of time" compared to smaller, uncompressed models. The central controversy is whether the significant quality degradation of a large model is a worthwhile compromise for accessibility, with many finding 4-bit quantization to be a more reliable minimum.

Project Nomad – Knowledge That Never Goes Offline

▲ 521 · 190 comments · projectnomad.us

TL;DR: Project NOMAD is a free, open-source server providing offline access to Wikipedia, AI, maps, and education, enabling digital independence for emergencies and off-grid scenarios.

Project NOMAD, a free and open-source offline server, enables users to access knowledge, AI, maps, and educational tools without an internet connection on their own hardware. It integrates technologies like Kiwix for offline encyclopedias, Ollama for local large language models, OpenStreetMap for navigation, and Kolibri for educational content including Khan Academy courses. Unlike other solutions that can cost hundreds, NOMAD is free and is built for robust systems, supporting GPU-accelerated AI on

recommended specs such as AMD Ryzen 7 or Intel i7+ processors with 32GB RAM. This makes it ideal for emergency preparedness, off-grid living, and tech enthusiasts seeking digital independence and full data control.

WHAT THE COMMUNITY SAYS

The comments center on a debate over a project's technical flaws and its "prepper" branding. Critics point out practical issues like its US-centric focus and a broken Docker setup, while others dismiss the military aesthetic as "larping." Conversely, defenders find the project's goal of preserving offline knowledge valuable, especially as a countermeasure to internet shutdowns by authoritarian regimes, arguing that its utility outweighs any aesthetic concerns.

The future of version control

▲ 589 · 330 comments · bramcohen.com

TL;DR: Manyana presents a vision for the future of version control by using CRDTs to ensure merges always succeed, offering informative conflict presentation instead of blocking traditional conflicts.

A new project called Manyana introduces a compelling vision for the future of version control, leveraging Conflict-Free Replicated Data Types (CRDTs). This system fundamentally redefines conflict resolution, ensuring merges always succeed while providing significantly more informative conflict markers that detail "what happened" and "who did it," rather than opaque code blocks. CRDTs enable eventual consistency, guaranteeing the same merge result regardless of branch order, and allow for non-blocking conflict reviews. Notably, Manyana also offers a unique approach to rebase that preserves full history.

WHAT THE COMMUNITY SAYS

The comments focus on the user experience of Git's merge conflict resolution, with a consensus that the default presentation is confusing. The main debate is whether to solve this with external tools like `p4merge` or by using built-in Git configurations like `diff3` to add more context. A key insight is that even with technical solutions, the fundamental vocabulary of Git (e.g., "ours" vs. "theirs", "HEAD") remains a significant and long-standing source of confusion for many developers.

PC Gamer recommends RSS readers in a 37mb article that just keeps downloading

▲ 725 · 339 comments · stuartbreckenridge.net

TL;DR: A PC Gamer article recommending RSS readers ironically exemplifies excessive web bloat with 37MB initial load, numerous ads, and continuous downloads, underscoring the efficiency of RSS.

A recent PC Gamer article recommending RSS readers has sparked criticism for its own poor web practices. Upon navigating to the page, users are met with multiple intrusive pop-ups and at least five visible ads, obscuring the content. The article itself is a hefty 37MB on initial load, and alarmingly, continuously downloads new advertisements, consuming nearly half a gigabyte of data within just five minutes. This highlights the utility of RSS readers like NetNewsWire and Reeder in bypassing such ad-heavy and data-intensive online experiences.

WHAT THE COMMUNITY SAYS

The comments primarily criticize the excessive data consumption of modern websites, exemplified by one site downloading 500MB in minutes due to ads. The main debate revolves around this "blatant negligence of frugality," where bloated software requires immense resources for simple tasks, disproportionately harming users on metered data plans, such as those on government assistance who are quickly locked out of essential services. A key insight is the contrast between a broken, heavy website and an efficient, low-bandwidth alternative like SMS accomplishing the same goal, highlighting a broader frustration with web development trends.

Reports of code's death are greatly exaggerated

▲ 500 · 351 comments · stevekrouse.com

TL;DR: While AI-driven "vibe coding" from natural language specifications seems easy, it creates an illusion of precision that cannot replace building robust abstractions to manage hidden software complexities.

A recent analysis argues that while AI makes turning natural language into code easier, reports of code's death are "greatly exaggerated." The author warns of "vibe coding," where developers use high-level prompts that feel precise but are not, leading to "leaky

abstractions" and unexpected bugs when systems scale. For example, a vibe-coded app with "live collaboration" went viral and then crashed because the developer underestimated the feature's immense underlying complexity. The essay concludes that the fundamental skill of programmers is not writing code, but mastering complexity through precise abstractions, a need that AI does not eliminate.

WHAT THE COMMUNITY SAYS

The discussion centers on whether AI can truly innovate or is fundamentally limited to regurgitating existing knowledge. The main perspective argues that AI is a "conformist," trained on vast datasets of human work, which causes it to align with consensus and struggle with the independent, critical thinking required to advance the state of the art. A complementary view finds value in this, noting that AI excels at accelerating tedious but non-innovative tasks like system integration and serves as a powerful "rubber duck" for developers to refine their own thoughts. The core tension is whether AI's current inability to create genuine, paradigm-shifting breakthroughs is a temporary hurdle or a fundamental constraint of the technology.

GrapheneOS will remain usable by anyone without requiring personal information

▲ 530 · 150 comments · grapheneos.social

TL;DR: GrapheneOS affirms its commitment to privacy, ensuring usability without personal information by leveraging AOSP and partnerships to bypass Google dependencies and resist anti-FOSS efforts.

GrapheneOS reaffirms its commitment to user privacy, allowing full functionality without requiring personal information. The operating system is increasingly challenged by difficulties in supporting Pixel devices, making its partnership with Motorola critical for continued development. Despite its ties to Google, the underlying Android Open Source Project (AOSP) remains viable for creating mobile OS experiences independent of Google services. The article also highlights that Free and Open-Source Software (FOSS), like encryption, is resilient against attempts to ban it, though it often faces efforts to discredit it or impose backdoors from authoritarian entities.

WHAT THE COMMUNITY SAYS

The discussion centers on Android's geographic restrictions, primarily the disabling of call recording due to varied US consent laws, which some users see as a user-hostile policy that empowers corporations. The key debate is whether this broad approach is a necessary legal precaution or an overreach, with users suggesting technical solutions for more granular control. A related insight is that other regional features, like Japan's mandatory camera shutter sound, may not be legal mandates but rather industry self-regulation to preempt legislation and address significant social issues.

Building an FPGA 3dfx Voodoo with Modern RTL Tools

▲ 216 · 48 comments · noquiche.fyi

TL;DR: An FPGA reimplementation of the classic 3dfx Voodoo 1 graphics card was successfully built using modern RTL tools, demonstrating that complex hardware can now be accurately recreated and debugged by individuals.

A developer successfully rebuilt a 3dfx Voodoo 1 graphics card using an FPGA and modern RTL tools like SpinalHDL, demonstrating that complex designs can now be individually described, simulated, and debugged. The project, available on GitHub, highlights the Voodoo 1's unique complexity stemming from its hardwired fixed-function rendering behaviors, such as Gouraud shading and trilinear filtering, rather than modern programmable shaders. A key challenge involved precisely replicating the original Voodoo's behavior, exemplified by a bug where translucent pixels resulted from multiple subtle hardware-accuracy mismatches. This was made tractable by innovative approaches to representing register semantics and debugging deep graphics pipelines.

WHAT THE COMMUNITY SAYS

The comments express nostalgia for the Voodoo graphics cards' groundbreaking performance and aesthetics for their time. A key debate arises comparing Voodoo's practical triangle-based rendering to the Nvidia NV-1's more ambitious but difficult-to-program NURBS-based system, with the consensus that Voodoo's simplicity rightfully won. Other insights include an appreciation for the alternative

"pixellated" charm of software rendering and the desire to frame the physical cards as nostalgic artifacts.