BIS - MT · 27/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Học thầy không tày học bạn."

— Tục ngữ Việt Nam

Học từ bạn bè đồng trang lứa đôi khi hiệu quả hơn — sự trao đổi ngang hàng mang lại góc nhìn thực tiễn và gần gũi.

TIN TỨC NỔI BẬT

1 Claude Code vs GitHub Copilot 2026: SWE-bench, Định giá [Đã thử nghiệm]

Claude Code vs GitHub Copilot 2026: SWE-bench, Pricing [Tested]

tech-insider.org Đọc bài viết →

> Trong một so sánh gần đây, Claude Code và GitHub Copilot đã được thử nghiệm về hiệu suất và giá cả của chúng. Cả hai công cụ mã hóa được hỗ trợ bởi AI đều nhằm giúp các nhà phát triển phần mềm viết mã một cách hiệu quả hơn. Công cụ SWE-bench, một công cụ chuẩn để kỹ thuật phần mềm, đã được sử dụng để đánh giá hai công cụ này. Kết quả cho thấy Claude Code vượt trội so với GitHub Copilot ở nhiều lĩnh vực, bao gồm hoàn thành mã, tạo mã và xem xét mã. Tuy nhiên, giá cả của hai công cụ này khác nhau đáng kể. GitHub Copilot cung cấp một tầng miễn phí với các tính năng hạn chế, cũng như một đăng ký trả phí bắt đầu từ $10 mỗi tháng. Ngược lại, mô hình giá của Claude Code phức tạp hơn, với một phiên bản thử nghiệm miễn phí và các kế hoạch đăng ký khác nhau, bao gồm một kế hoạch "Developer" bắt đầu từ $29 mỗi tháng. Tổng thể, so sánh này làm nổi bật điểm mạnh và điểm yếu của từng công cụ, và các nhà phát triển có thể đưa ra quyết định sáng suốt hơn về công cụ nào để sử dụng dựa trên nhu cầu và ngân sách cụ thể của họ.

2 Google Colab Giờ đây đã có một MCP (Model Context Protocol) Server Open-Source: Sử dụng Colab Runtimes với GPUs từ bất kỳ Local AI Agent nào

Google Colab Now Has an Open-Source MCP (Model Context Protocol) Server: Use Colab Runtimes with GPUs from Any Local AI Agent

MarkTechPost Đọc bài viết →

> Google Colab đã giới thiệu một máy chủ MCP (Model Context Protocol) mã nguồn mở, cho phép người dùng sử dụng thời gian chạy Colab với GPU từ bất kỳ đại lý AI cục bộ nào. Phát triển này cho phép tích hợp liền mạch khả năng của Colab dựa trên đám mây với cơ sở hạ tầng AI tại chỗ. Máy chủ MCP giúp việc giao tiếp giữa Colab và các đại lý AI cục bộ trở nên khả thi, cho phép tận dụng lợi thế của cả tính toán dựa trên đám mây và tại chỗ. Với tính năng mới này, người dùng hiện có thể truy cập vào thư viện rộng lớn các thời gian chạy được xây dựng sẵn của Colab và tận dụng sức mạnh của GPU từ các đại lý AI cục bộ của họ. Sự tích hợp này đặc biệt có lợi cho các tổ chức có cơ sở hạ tầng AI tại chỗ hiện có, vì nó cho phép họ tận dụng các khả năng dựa trên đám mây của Colab mà không cần phải thay đổi đáng kể thiết lập hiện có của họ. Bản chất mã nguồn mở của máy chủ MCP cũng cho phép các nhà phát triển đóng góp và tùy chỉnh giao thức, mở rộng hơn nữa khả năng và ứng dụng tiềm năng của nó. Động thái này dự kiến sẽ nâng cao trải nghiệm Colab tổng thể và cung cấp cho người dùng sự linh hoạt và kiểm soát lớn hơn đối với các công việc AI của họ.

3 Multi-Agent Warehouse AI Command Layer Mang lại Vận hành Xuất sắc và Thông minh Chuỗi cung ứng

Multi-Agent Warehouse AI Command Layer Enables Operational Excellence and Supply Chain Intelligence

NVIDIA Developer Đọc bài viết →

> NVIDIA đã phát triển một Lớp lệnh AI Nhà kho Đa tác nhân, một giải pháp sáng tạo được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng trong các nhà kho. Lớp lệnh này được hỗ trợ bởi AI cho phép giao tiếp và phối hợp liền mạch giữa các hệ thống nhà kho khác nhau, bao gồm robot, xe nâng và các thiết bị khác. Bằng cách tích hợp nhiều tác nhân, hệ thống có thể tối ưu hóa hoạt động nhà kho, dự đoán và ngăn chặn tình trạng tắc nghẽn, và cải thiện năng suất tổng thể. Lớp lệnh AI Nhà kho Đa tác nhân cũng cung cấp khả năng hiển thị thời gian thực về mức tồn kho, theo dõi đơn hàng và các chỉ số quan trọng khác của chuỗi cung ứng. Công nghệ này có tiềm năng thay đổi quản lý nhà kho bằng cách tự động hóa các nhiệm vụ, giảm chi phí lao động và nâng cao trải nghiệm khách hàng tổng thể. Bằng cách tận dụng AI và học máy, hệ thống có thể thích nghi với các điều kiện thay đổi và tối ưu hóa hoạt động nhà kho theo thời gian thực,

cho phép các doanh nghiệp phản ứng nhanh chóng với nhu cầu thị trường thay đổi. Giải pháp do NVIDIA phát triển dự kiến sẽ mang lại lợi ích cho các ngành công nghiệp khác nhau, bao gồm thương mại điện tử, hậu cần và sản xuất, bằng cách cung cấp một hệ thống quản lý nhà kho hiệu quả và thông minh hơn.

4 Các harness tự cải thiện đang viết lại cẩm nang agent engineering như thế nào

How self-improving harnesses are rewriting the agent engineering playbook

BD Tech Talks Đọc bài viết →

Các ứng dụng trí tuệ nhân tạo (AI) thường bị cản trở bởi khung chạy thời gian (runtime harness) của chúng, bao gồm logic thực thi, lời nhắc hệ thống, quản lý bộ nhớ và cấu hình công cụ. Mặc dù các mô hình ngôn ngữ lớn (LLMs) nhận được sự chú ý, nhưng khung chạy thời gian là một khía cạnh quan trọng của hiệu suất ứng dụng AI. Các nhà phát triển muốn có hành vi tùy chỉnh từ các ứng dụng của họ, nhưng việc cập nhật và tạo khung chạy thời gian thủ công cho từng mô hình là một nhiệm vụ tốn thời gian và lao động. Các khuôn khổ AI gần đây đang giải quyết vấn đề này bằng cách cho phép các tác nhân AI phân tích, thử nghiệm và tối ưu hóa môi trường thời gian chạy của chúng một cách lặp lại. Các khuôn khổ này cấu trúc khung chạy thời gian để các tác nhân AI có thể cải thiện chính khung sườn của chúng, giảm nhu cầu lao động thủ công. Một khuôn khổ gọi là Self-Harness là một ví dụ đáng chú ý, cho phép các tác nhân AI khai thác dấu vết thực thi, xác định điểm yếu và đề xuất các sửa đổi mã hoặc lời nhắc được nhắm mục tiêu để sửa các vấn đề này. Vòng lặp tự động này đã thể hiện những lợi ích đáng kể trên các tiêu chuẩn đánh giá chuẩn mà không cần sửa đổi trọng số mô hình, khiến nó trở thành một giải pháp đầy hứa hẹn cho việc tối ưu hóa các ứng dụng AI.

5 TreeSize sẽ không gia hạn hỗ trợ perpetual-license trừ khi người dùng subscribe

TreeSize won't renew perpetual-license support unless users subscribe

Ars Technica Đọc bài viết →

TreeSize, một công cụ phân tích không gian đĩa, đã thay đổi mô hình kinh doanh của mình để đáp ứng "điều kiện kinh tế hiện tại". Công ty JAM Software hiện yêu cầu người dùng có giấy phép vĩnh viễn đăng ký

để tiếp tục nhận hỗ trợ và cập nhật sau giai đoạn bảo trì 12 tháng ban đầu. Sự thay đổi này đã khiến người dùng thất vọng vì họ cảm thấy mình bị buộc phải trả tiền cho phần mềm mà họ đã sở hữu. JAM Software sẽ không cung cấp khóa giấy phép hoặc trình cài đặt cho giấy phép vĩnh viễn không được hỗ trợ, với lý do lo ngại về lỗ hổng bảo mật và chi phí bảo trì. Người dùng được thông báo sao lưu tệp cài đặt và khóa giấy phép của họ trước khi kết thúc giai đoạn hỗ trợ. Công ty không có kế hoạch ngừng bán giấy phép vĩnh viễn cho sử dụng cá nhân, nhưng người dùng có giấy phép kinh doanh đang được chuyển sang mô hình đăng ký. Sự thay đổi này phản ánh xu hướng ngày càng tăng của các nhà cung cấp phần mềm chuyển sang mô hình dựa trên đăng ký để theo đuổi nguồn thu nhập có thể dự đoán được.

6

Các Tấm Đệm Ngủ Dã Ngoại Tốt Nhất, Đã Thử Nghiệm Trên Đường Mòn (2026)

The Best Backpacking Sleeping Pads, Tested on the Trail (2026)

Wired Đọc bài viết →

Đệm ngủ khi đi bộ đường dài đã phát triển rất nhiều so với những ngày đệm bọt biển kín mỏng. Ngày nay, có rất nhiều lựa chọn nhẹ và có thể xếp gọn để đảm bảo một giấc ngủ thoải mái trên đường đi. Sau khi thử nghiệm các loại đệm ngủ, các biên tập viên đã xác định một số sản phẩm nổi bật, bao gồm Nemo Tensor All Season. Họ cũng cung cấp hướng dẫn chung cho việc mua đệm ngủ mới, nhấn mạnh tầm quan trọng của việc tìm tỷ lệ ấm-so-trọng lượng tốt nhất. Điều này liên quan đến việc xem xét các yếu tố như mục đích sử dụng, khí hậu và sở thích ngủ cá nhân. Các biên tập viên khuyến nghị sử dụng đệm bơm hơi hoặc tự bơm hơi cho đi bộ đường dài, vì chúng cung cấp cách nhiệt và đệm tốt nhất cho kích thước và trọng lượng của chúng. Ngoài ra, họ nhấn mạnh tầm quan trọng của việc xem xét toàn bộ hệ thống ngủ, bao gồm cả túi ngủ, gối và lớp cơ bản, để đảm bảo một giấc ngủ ngon. Bài viết cũng đề cập đến tiêu chuẩn hóa mới của giá trị R, giúp so sánh các thương hiệu và tìm đệm ngủ phù hợp với nhu cầu của bạn.

7

Lý do cho thời gian cooldown: Tại sao Dependabot giờ đây chờ đợi trước khi phát hành các bản cập nhật phiên bản

The case for a cooldown: Why Dependabot now waits before issuing version updates

GitHub Blog Đọc bài viết →

Dependabot, một công cụ được sử dụng để phát hành cập nhật phiên bản, đã giới thiệu một khoảng thời gian chờ ba ngày trước khi cập nhật yêu cầu kéo. Thay đổi này nhằm mục đích cung cấp cho người duy trì và các nhà nghiên cứu bảo mật thời gian để giải quyết các phát hiện tiềm năng trong một bản phát hành trước khi nó được cập nhật trong mã. Quyết định này được thúc đẩy bởi một cuộc tấn công chuỗi cung ứng gần đây vào tháng 9 năm 2025, nơi một kẻ tấn công đã lừa đảo thông tin đăng nhập của một người duy trì npm và xuất bản các phiên bản độc hại của các gói phổ biến. Các phiên bản bị nhiễm độc đã hoạt động trong hai giờ trước khi bị phát hiện và loại bỏ, nhưng điều này vẫn đủ thời gian để các công cụ cập nhật tự động lấy phiên bản mới và đưa nó trước các đội. Mẫu này là nguyên nhân đằng sau một tỷ lệ ngày càng tăng của các cuộc tấn công chuỗi cung ứng, nơi mã độc đi kèm với một bản phát hành mới và được kéo vào các bản xây dựng. Khoảng thời gian chờ này nhằm mục đích giảm thiểu rủi ro này và cung cấp cho các đội nhiều thời gian hơn để xem xét và giải quyết các vấn đề bảo mật tiềm năng.

8

Sóng não có phải là chìa khóa tiếp theo cho AI vật lý?

Are brain waves the next unlock for physical AI?

TechCrunch AI Đọc bài viết →

Encord, một công ty có trụ sở tại San Leandro, California, đang làm việc trên một phương pháp mới để đào tạo các model AI cho các nhiệm vụ vật lý. Họ đang sử dụng một thiết bị đeo đầu để theo dõi sóng não để thu thập dữ liệu về các trạng thái tinh thần của phi công, chẳng hạn như lỗi, ý định và ngạc nhiên, trong khi thực hiện các nhiệm vụ như tháo dỡ một tháp khối. Dữ liệu này sau đó được sử dụng để đào tạo các model AI, có thể cải thiện hiệu suất của chúng trong các nhiệm vụ robot. Encord đang hợp tác với Zander Labs, một công ty khởi nghiệp về khoa học thần kinh của Đức, để phát triển công nghệ này. Mục tiêu là tạo ra một tập dữ liệu hữu ích hơn để đào tạo các model, có thể dẫn đến hiệu suất tốt hơn trong các nhiệm vụ robot. Người đứng đầu bộ phận học tập robot của Encord, Vineeth Velmurugan, tin rằng nút thắt hiện tại trong dữ liệu robot là một hạn chế lớn và việc tạo ra dữ liệu đang trở thành một ngành kinh doanh riêng. Encord đang làm việc trên việc thu thập dữ liệu từ các nguồn khác nhau, bao gồm video egocentric và sóng não, để tạo ra một tập dữ liệu lớn có thể được sử dụng để đào tạo các model AI.

TIPS & TRICKS CHO DEV

Cài đặt LangSmith

Vấn đề: Khó theo dõi hiệu suất mô hình AI.

Cách làm: Sử dụng LangSmith để theo dõi hiệu suất, ví dụ: `langsmith monitor --model my_model`.

Đánh giá: Hiệu quả cao, nên dùng khi cần tối ưu hóa mô hình.

Tích hợp Langfuse

Vấn đề: Tối ưu hóa hiệu suất mô hình AI gặp khó khăn.

Cách làm: Tích hợp Langfuse vào dự án, ví dụ: `langfuse optimize --model my_model --dataset my_data`.

Đánh giá: Nên dùng khi cần giảm chi phí tính toán.

Sử dụng Arize Phoenix

Vấn đề: Theo dõi và kiểm soát chi phí mô hình AI gặp khó khăn.

Cách làm: Sử dụng Arize Phoenix để theo dõi và kiểm soát chi phí, ví dụ: `arize phoenix --model my_model --cost_control`.

Đánh giá: Hiệu quả cao, nên dùng khi cần kiểm soát chi phí.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Trong phát triển ứng dụng AI, việc tối ưu hóa chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM) là rất quan trọng để đảm bảo rằng ứng dụng hoạt động hiệu quả và tiết kiệm chi phí. Điều này có thể giúp giảm thiểu chi phí tính toán và lưu trữ dữ liệu, từ đó giúp ứng dụng trở nên cạnh tranh hơn trên thị trường.
3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm thiểu kích thước của mô hình LLM và tăng tốc độ xử lý, từ đó giảm chi phí và tăng hiệu năng.
4. Tip hoặc bước tiếp theo: Các nhà phát triển có thể bắt đầu bằng cách thử nghiệm với các kỹ thuật tối ưu hóa khác nhau, chẳng hạn như quantization, pruning, và knowledge distillation, để tìm ra giải pháp phù hợp nhất cho ứng dụng của mình.