BIS - MT · 02/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Lửa thử vàng, gian nan thử sức."

— Tục ngữ Việt Nam

Khó khăn là phép thử thực sự của ý chí — chỉ khi đối mặt thử thách, ta mới khám phá được năng lực thật sự của mình.

TIN TỨC NỔI BẬT

1

Claude Code vs GitHub Copilot 2026: SWE-bench, 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 kiểm tra về hiệu suất và giá cả của chúng. Công cụ SWE-bench, một công cụ dùng để đánh giá hiệu quả của các công cụ phát triển phần mềm, đã được sử dụng để đánh giá khả năng mã hóa của cả hai trợ lý AI. Kết quả cho thấy Claude Code vượt trội so với GitHub Copilot ở nhiều lĩnh vực, bao gồm chất lượng mã, tốc độ và độ chính xác. Tuy nhiên, mô hình giá của GitHub Copilot cung cấp một tầng miễn phí, cho phép người dùng truy cập vào các tính năng của nó mà không phải chịu chi phí. Ngược lại, Claude Code yêu cầu đăng ký, với các kế hoạch giá từ $15 đến $50 mỗi tháng. So sánh này nhấn mạnh sự đánh đổi giữa hiệu suất và chi phí khi lựa chọn giữa hai công cụ mã hóa AI này. Trong khi Claude Code cung cấp khả năng mã hóa vượt trội, giá của nó có thể là một rào cản đối với một số người dùng. Mặt khác, tầng miễn phí của GitHub Copilot khiến nó trở thành một lựa chọn dễ tiếp cận hơn cho các developer, mặc dù hiệu suất của nó có thể không mạnh như Claude Code.

2

Lớp lệnh AI Multi-Agent cho kho hàng: Đạt hiệu quả vận hành vượt trội 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 giới thiệu Lớp lệnh AI Nhà kho Đa tác nhân, được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng trong các nhà kho. Giải pháp này được hỗ trợ bởi AI cho phép giao tiếp và phối hợp liền mạch giữa các hệ thống khác nhau, bao gồm robot, xe nâng và các thiết bị khác. Bằng cách tận dụng các thuật toán AI và học máy tiên tiến, lớp lệnh có thể tối ưu hóa hoạt động nhà kho, dự đoán và ngăn chặn các nút thắt, và cải thiện năng suất tổng thể. Kiến trúc đa tác nhân của hệ thống cho phép nó thích nghi với các điều kiện nhà kho thay đổi và đưa ra quyết định theo thời gian thực. Điều này cho phép các nhà kho phản ứng nhanh với nhu cầu thay đổi và tối ưu hóa quy trình làm việc của họ cho phù hợp. Ngoài ra, lớp lệnh cung cấp thông tin chi tiết quý giá về hoạt động chuỗi cung ứng, cho phép các doanh nghiệp đưa ra quyết định dựa trên dữ liệu và cải thiện khả năng chống chịu chuỗi cung ứng tổng thể của họ. Bằng cách triển khai Lớp lệnh AI Nhà kho Đa tác nhân, các nhà kho có thể đạt được sự xuất sắc trong hoạt động, giảm chi phí và nâng cao khả năng đáp ứng nhu cầu của khách hàng. Giải pháp này có tiềm năng cách mạng hóa cách các nhà kho hoạt động, làm cho chúng trở nên hiệu quả, linh hoạt và phản ứng nhanh với các điều kiện thị trường thay đổi.

3

Vai trò thầm lặng của Toán học và Algorithms trong MCP & Multi-Agent Systems

The Silent Role of Mathematics and Algorithms in MCP & Multi-Agent Systems

Cisco Blogs Đọc bài viết →

Bài viết "Vai trò im lặng của Toán học và Thuật toán trong MCP & Hệ thống Đa tác nhân" nhấn mạnh sự đóng góp quan trọng nhưng thường bị bỏ qua của toán học và thuật toán trong các hệ thống phức tạp. Hệ thống Đa tác nhân (MAS) và Nền tảng Đa đám mây (MCP) phụ thuộc nặng nề vào các mô hình toán học và thuật toán để hoạt động hiệu quả. Trong MAS, các mô hình toán học được sử dụng để mô phỏng và phân tích các tương tác phức tạp giữa nhiều tác nhân, cho phép các nhà nghiên cứu hiểu và dự đoán hành vi của chúng. Các mô hình này cũng giúp tạo ra các thuật toán ra quyết định tối ưu hóa kết quả trong các môi trường động. Tương tự, MCP dựa vào các thuật toán để quản lý và điều phối tài nguyên trên nhiều nền tảng đám mây, đảm bảo tích hợp liền mạch và phân bổ tài nguyên hiệu quả. Các thuật toán này cho phép MCP cung cấp các dịch vụ có thể mở rộng, bảo mật và đáng tin cậy cho người dùng. Bài viết nhấn mạnh tầm quan trọng của toán học và thuật toán trong việc cho phép phát triển và vận hành các hệ thống

phức tạp như MAS và MCP. Bằng cách tận dụng các mô hình toán học và thuật toán, các nhà nghiên cứu và nhà phát triển có thể tạo ra các hệ thống hiệu quả, bảo mật và đáng tin cậy hơn, đáp ứng nhu cầu của các ứng dụng hiện đại.

4

Ưu đãi Grill và Griddle tốt nhất dịp 4 tháng 7: Weber, Traeger, Recteq

The Best July 4 Grill and Griddle Deals: Weber, Traeger, Recteq

Wired Đọc bài viết →

Khi ngày 4 tháng 7 đang đến gần, đây là dịp cuối cùng để có được những ưu đãi tốt nhất cho nướng BBQ trong suốt mùa hè. Những lựa chọn hàng đầu từ WIRED bao gồm Recteq Flagship 1,600 pellet smoker, hiện đang giảm giá 250 đô la. Thiết bị nướng này có khả năng kết nối Wi-Fi, có diện tích nấu 1,600 inch vuông và có thùng chứa 40 pound cho các món ăn dài và nhiệt độ đều. Thiết bị griddle Traeger Flatrock cũng được khuyến nghị cao, với bề mặt dẫn nhiệt và thiết kế bộ đốt hình U sáng tạo cho phân phối nhiệt hoàn hảo. Phiên bản 2 bộ đốt của Traeger Flatrock hiện giảm giá 100 đô la, với giá 699 đô la vào cuối tuần ngày 4 tháng 7. Ngoài ra, dòng griddle Traeger Irontop cung cấp một lựa chọn tiết kiệm hơn, với mô hình 4 bộ đốt có sẵn với giá 599 đô la. Các ưu đãi khác bao gồm nướng khí Weber Spirit 200 series, hiện giảm giá 50 đô la, và dòng griddle Weber Slate đã được tẩm ướp trước, cũng giảm giá 50 đô la. Nướng pellet Traeger Woodridge Pro là một lựa chọn hàng đầu khác, hiện giảm giá 150 đô la và có sẵn với giá dưới 1.000 đô la. Những ưu đãi này chỉ có sẵn vào cuối tuần ngày 4 tháng 7, vì vậy hãy hành động nhanh để tận dụng những tiết kiệm này.

5

Google tạo ra smart speaker tuyệt vời, nhưng Gemini chưa sẵn sàng

Google built a great smart speaker, but Gemini isn't ready for it

The Verge AI Đọc bài viết →

Google đã phát hành loa thông minh mới đầu tiên trong sáu năm, Loa Google Home, được xây dựng cụ thể cho trợ lý Gemini. Phần cứng của loa đã nhận được lời khen ngợi về thiết kế và chất lượng âm thanh, khiến nó trở thành một bổ sung tuyệt vời cho bất kỳ phòng nào. Nó nhỏ gọn, hấp dẫn và phải chăng, với màu ngọc bích mềm mà hòa lẫn

hoàn hảo. Loa có âm thanh 360 độ và có thể kết hợp với các loa khác để tạo hiệu ứng âm thanh nổi. Tuy nhiên, trợ lý Gemini, điều khiển loa, vẫn cảm giác chưa hoàn thiện. Nó chậm và không đáng tin cậy đôi khi, và một số tính năng bị khóa sau thanh toán. Ngoài ra, chất lượng âm thanh của loa không tốt như Nest Audio mà nó thay thế, đặc biệt là về âm trầm. Mặc dù những vấn đề này, Loa Google Home là một thiết bị vững chắc có thể là một lựa chọn tuyệt vời cho những người đang tìm kiếm một loa thông minh, nhưng nó có thể không phải là lựa chọn tốt nhất cho những người đã đầu tư mạnh vào hệ sinh thái nhà thông minh.

6

Cuộc đua vũ trang AI: Từ smart model đến hạ tầng full-stack

How the AI arms race moved from smart models to full-stack infrastructure

BD Tech Talks Đọc bài viết →

Cuộc đua vũ trang trí tuệ nhân tạo (AI) đã phát triển từ việc tập trung vào phát triển các mô hình thông minh thành một cuộc chiến cơ sở hạ tầng nhiều lớp. Bộ xếp chồng AI hiện đại bao gồm bốn lớp chính: phần cứng, cụm tính toán, lớp mô hình và lớp ứng dụng. Các công ty không còn hài lòng với việc thống trị một lớp duy nhất, vì điều này không còn là lợi thế lâu dài khả thi. Thay vào đó, các công ty thành công đang mở rộng sang các lớp liền kề để bảo vệ biên độ và nắm bắt giá trị mới. Để đạt được điều này, các công ty đang đầu tư vào cơ sở hạ tầng tùy chỉnh, chẳng hạn như chip suy luận AI tùy chỉnh và trung tâm dữ liệu chuyên dụng. Ví dụ, OpenAI đang hợp tác với Broadcom để ra mắt chip suy luận AI tùy chỉnh đầu tiên, trong khi Anthropic đã ký kết một quan hệ đối tác cơ sở hạ tầng 50 tỷ USD với nhà cung cấp neocloud FluidStack. Một số công ty, như xAI, thậm chí còn bỏ qua đám mây hoàn toàn bằng cách xây dựng siêu máy tính của riêng họ. Trong khi đó, các công ty lớp ứng dụng cũng đang mở rộng khả năng của mình bằng cách phát triển mô hình của riêng họ và tích lũy dữ liệu người dùng. Điều này cho phép họ xây dựng các rào cản phòng thủ và giảm sự phụ thuộc vào các API tiền phong tốn kém. Kết quả là, các công ty lớp ứng dụng đang trở nên hấp dẫn hơn đối với các gã khổng lồ tính toán lớp cơ sở, những người tuyệt vọng muốn đảm bảo các điểm chạm người dùng.

7

Trích dẫn từ Anthropic

Quoting Anthropic

Simon Willison Đọc bài viết →

> Bộ Thương mại đã đưa ra một thông báo quan trọng liên quan đến kiểm soát xuất khẩu đối với hai mô hình AI, Claude Fable 5 và Mythos 5. Theo thông báo, các kiểm soát xuất khẩu đã được dỡ bỏ, cho phép khôi phục quyền truy cập vào các mô hình này. Sự thay đổi này dự kiến sẽ có hiệu lực vào ngày mai, với các cập nhật thêm sẽ được chia sẻ trong tương lai gần. Tin tức này được thu thập bởi Simon Willison và đăng vào ngày 30 tháng 6 năm 2026. Việc dỡ bỏ kiểm soát xuất khẩu đối với các mô hình AI này cho thấy sự thay đổi trong các chính sách quản lý, có khả năng mở ra những cơ hội mới cho nghiên cứu, phát triển và triển khai các công nghệ này. Tuy nhiên, các ý nghĩa chính xác của sự thay đổi này vẫn còn phải được xem xét và có khả năng sẽ được làm rõ trong những ngày tới.

8

LLM mắc kẹt trong lối tư duy tập thể. Startup này đang tìm cách đưa chúng thoát ra.

LLMs are stuck in a groupthink groove. This startup is trying to get them out.

MIT Tech Review Đọc bài viết →

> Các mô hình ngôn ngữ lớn (LLMs) thường dễ đoán và ít sáng tạo trong các phản hồi của chúng hơn dự kiến, điều này có thể là một vấn đề khi suy nghĩ hoặc tìm kiếm ý tưởng độc đáo. Một trò chơi minh họa vấn đề này liên quan đến việc yêu cầu các LLM phổ biến, chẳng hạn như Claude và ChatGPT, tạo ra một số ngẫu nhiên giữa 1 và 10, kết quả là cùng một số, 7, hầu hết thời gian. Sự dự đoán này là do cách mà hầu hết các LLM được đào tạo trên dữ liệu tương tự để thực hiện các nhiệm vụ tương tự. Một công ty khởi nghiệp của Úc có tên Springboards đã phát triển một LLM gọi là Flint, nhằm phá vỡ khuôn mẫu này bằng cách tạo ra nhiều loại phản hồi đa dạng hơn cho các câu hỏi mở. Phản hồi của Flint đối với trò chơi số ngẫu nhiên và các lời nhắc khác, chẳng hạn như đặt tên cho một loại xe hơi, khác biệt đáng kể so với phản hồi của Claude và ChatGPT. Điều này là vì Flint đã được đào tạo để "chào đón ảo giác", những phản hồi sai hoặc không mong đợi có thể dẫn đến các câu trả lời sáng tạo và đa dạng hơn. Các nhà nghiên cứu cũng đã nhấn mạnh vấn đề về sự lặp lại trong các LLM, với một nghiên cứu gần đây cho thấy các mô hình khác nhau hội tụ trên các câu trả lời tương tự cho các câu hỏi mở. Giới hạn này đang bắt đầu

> nhận được sự chú ý nhiều hơn, với một số chuyên gia suy đoán rằng nó có thể là do các phương pháp đào tạo và dữ liệu tương tự được sử dụng để phát triển hầu hết các LLM.

TIPS & TRICKS CHO DEV

Orchestrate AI Agents

Vấn đề: Xử lý task phức tạp cần phối hợp nhiều AI agents.
Cách làm: Sử dụng LangGraph, CrewAI, AutoGen để tạo và quản lý các agents. Ví dụ: `langgraph init` để tạo dự án mới.
Đánh giá: Hiệu quả cao khi xử lý task phức tạp, nhưng cần kinh nghiệm về AI orchestration.

Deploy Multi-Agent Systems

Vấn đề: Triển khai hệ thống nhiều AI agents hiệu quả.
Cách làm: Sử dụng phiData để deploy và quản lý các agents. Ví dụ: `phidata deploy` để triển khai hệ thống.
Đánh giá: Hiệu quả cao khi cần deploy và quản lý nhiều agents, nhưng cần kiến thức về DevOps.

Monitor AI Agents

Vấn đề: Theo dõi và giám sát hiệu suất của các AI agents.
Cách làm: Sử dụng các công cụ như Prometheus và Grafana để monitor các agents. Ví dụ: `prometheus --help` để xem các tùy chọn.
Đánh giá: Hiệu quả cao khi cần theo dõi và giám sát hiệu suất của các agents, giúp tối ưu hóa hệ thống.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để xây dựng ứng dụng AI hiệu quả, các nhà phát triển cần tối ưu hóa chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM). Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý, đồng thời đảm bảo hiệu suất của mô hình.
3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và tăng tốc độ huấn luyện, từ đó giảm chi phí và tăng hiệu năng.
4. Tip: Để bắt đầu, hãy khám phá các kỹ thuật tối ưu hóa như quantization, pruning và knowledge distillation để giảm kích thước và tăng tốc độ của mô hình LLM, và áp dụng chúng vào dự án của bạn.

Luôn đi đầu trong thế giới AI! · Stay ahead in AI!

Nguồn: Google News · Groq AI