BIS - MT · 15/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

> *“Thời gian là vàng bạc.”*
>
> — Tục ngữ Việt Nam

Thời gian là tài sản quý nhất — một khi đã qua đi, không thể lấy lại, hãy sử dụng từng giây phút có ý nghĩa.

TIN TỨC NỔI BẬT

1

Claude Code vs GitHub Copilot 2026: SWE-bench, Giá cả [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 trong một điểm chuẩn Kỹ thuật Phần mềm (SWE). Kết quả cho thấy cả hai công cụ mã hóa được hỗ trợ bởi AI đều có điểm mạnh và điểm yếu. Claude Code, được phát triển bởi Anthropic, đã thể hiện hiệu suất vượt trội trong một số nhiệm vụ, chẳng hạn như hoàn thành mã và gỡ lỗi. Tuy nhiên, nó gặp khó khăn với các nhiệm vụ phức tạp hơn, như tái cấu trúc mã và tối ưu hóa. Mặt khác, GitHub Copilot đã xuất sắc trong tái cấu trúc mã và tối ưu hóa, nhưng không đạt được thành công trong hoàn thành mã và gỡ lỗi. Kiểm tra SWE-bench đã tiết lộ rằng hiệu suất của Claude Code đồng nhất hơn trên các nhiệm vụ khác nhau, trong khi hiệu suất của GitHub Copilot thay đổi tùy thuộc vào nhiệm vụ. Về giá cả, cả hai công cụ đều cung cấp mức giá cạnh tranh, với Claude Code bắt đầu từ 15 đô la mỗi tháng và GitHub Copilot có giá 10 đô la mỗi tháng cho cá nhân. Sự lựa chọn giữa hai công cụ cuối cùng phụ thuộc vào nhu cầu cụ thể của nhà phát triển và loại dự án họ làm việc.

2

Multi-Agent Warehouse AI Command Layer Nâng cao Hiệu quả Vận hành 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, một công nghệ tiên tiến được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng trong các nhà kho. Giải pháp đổi mới này tận dụng trí tuệ nhân tạo (AI) để tối ưu hóa hoạt động nhà kho, đảm bảo quản lý hàng tồn kho không gián đoạn, thực hiện đơn hàng hiệu quả và hậu cần được sắp xếp hợp lý. Lớp lệnh AI Nhà kho Đa tác nhân sử dụng kiến trúc phân tán, cho phép nhiều tác nhân AI làm việc cùng nhau để đưa ra quyết định thông minh trong thời gian thực. Điều này cho phép các nhà kho thích nghi nhanh với nhu cầu thay đổi, tối ưu hóa phân bổ tài nguyên và giảm thiểu lỗi. Lợi ích chính của công nghệ này bao gồm hiệu quả hoạt động được cải thiện, khả năng hiển thị chuỗi cung ứng được tăng cường và độ chính xác trong quản lý hàng tồn kho được tăng lên. Bằng cách tự động hóa các nhiệm vụ thường xuyên và cung cấp thông tin theo thời gian thực, Lớp lệnh AI Nhà kho Đa tác nhân cho phép các nhà kho đưa ra quyết định dựa trên dữ liệu, cuối cùng thúc đẩy sự phát triển và cạnh tranh của doanh nghiệp. Giải pháp được NVIDIA phát triển có khả năng cách mạng hóa phong cảnh quản lý nhà kho, cung cấp một nền tảng có thể mở rộng và linh hoạt cho các nhà kho để đạt được sự xuất sắc về hoạt động và trí tuệ chuỗi cung ứng.

3 AWS Open-Source MCP Server cho Bedrock AgentCore để Tối ưu Phát triển AI Agent

AWS Open-Sources an MCP Server for Bedrock AgentCore to Streamline AI Agent Development

MarkTechPost Đọc bài viết →

AWS đã thực hiện một bước tiến quan trọng trong không gian phát triển AI bằng cách mở nguồn một máy chủ MCP (Nền tảng Đa đám mây) cho Bedrock AgentCore. Động thái này nhằm mục đích đơn giản hóa việc phát triển các tác nhân AI và làm cho nó trở nên dễ tiếp cận hơn với các nhà phát triển. Máy chủ MCP được thiết kế để hoạt động với Bedrock AgentCore, một framework để xây dựng các tác nhân AI. Bằng cách mở nguồn máy chủ MCP, AWS đang cung cấp cho các nhà phát triển một nền tảng tiêu chuẩn hóa để triển khai và quản lý các tác nhân AI trên các môi trường đám mây khác nhau. Điều này có thể giúp giảm thiểu sự phức tạp và chi phí liên quan đến việc phát triển và triển khai các tác nhân AI. Việc mở nguồn máy chủ MCP dự kiến sẽ đẩy nhanh việc phát triển các tác nhân AI và làm cho nó dễ dàng hơn cho các nhà phát triển tích hợp chúng vào các ứng dụng của mình. Với động thái này, AWS đang mở rộng hơn nữa các dịch vụ của mình trong

không gian AI và học máy, và cung cấp cho các nhà phát triển nhiều công cụ và tài nguyên hơn để xây dựng các ứng dụng sáng tạo.

4 The Download: Cơ chế hoạt động bên trong của Claude và tương lai của world model

The Download: Claude's inner workings, and the future of world models

MIT Tech Review Đọc bài viết →

New York đã trở thành tiểu bang đầu tiên ban hành lệnh tạm ngừng xây dựng trung tâm dữ liệu, cấm xây dựng trung tâm dữ liệu lớn trong thời gian lên đến một năm. Động thái này diễn ra khi việc giao hàng điện thoại thông minh đã giảm xuống mức thấp nhất trong 13 năm do khan hiếm bộ nhớ, khiến giá tăng lên như một kết quả. Trong khi đó, các nhà nghiên cứu đã có một khám phá đột phá, tìm thấy các phân tử đường trong không gian liên sao, điều này có thể gợi ý về khả năng sự sống trên Trái Đất được gieo trồng từ không gian. Trong tin tức về AI, Anthropic đã công bố một cửa sổ mới vào "suy nghĩ nội bộ" của các model khi chúng suy luận thông qua các câu trả lời. Biên tập viên cao cấp Will Douglas Heaven đã nói về những ý nghĩa của nghiên cứu này, có thể dẫn đến sự phát triển của các model thế giới có thể bắc cầu giữa các hệ thống AI và thế giới vật lý. Một sự kiện LinkedIn Live hôm nay sẽ khám phá cách các model thế giới có thể biến đổi robot và mở khóa các máy móc thông minh. Các câu chuyện đáng chú ý khác bao gồm Nvidia giảm một nửa danh sách người mua châu Á để ngăn chặn các chip AI đến Trung Quốc, và các hacker nhà nước Nga nhắm vào các bộ định tuyến để gián điệp và đánh cắp. Ngoài ra, một liệu pháp tế bào mới đã cứu bốn đứa trẻ bị ung thư não giai đoạn cuối, và LAPD đã ngừng sử dụng camera giám sát Flock do lo ngại về quyền riêng tư.

5 Cách Debug Coding Agent với LangSmith Traces

How to Debug Coding Agents with LangSmith Traces

LangChain Blog Đọc bài viết →

Các nhà phát triển thường gặp khó khăn khi gỡ lỗi các tác nhân mã hóa do sự phức tạp và phân mảnh của các công cụ khác nhau và các cơ chế ghi lại của chúng. Mỗi tác nhân mã hóa, chẳng hạn như Claude, Codex hoặc Cursor, có cách riêng để ghi lại các bước trung gian và đầu ra, khiến việc quan sát và hiểu các quy trình ra quyết định của chúng trở nên thách thức. Điều này dẫn đến một quy trình gỡ lỗi thay

đổi mỗi khi nhà phát triển chuyển đổi công cụ, yêu cầu họ phải kiểm tra nhiều nơi và duy trì nhiều mô hình tinh thần cho các định dạng sự kiện khác nhau. Để giải quyết vấn đề này, LangSmith cung cấp một lớp quan sát thống nhất trên các tác nhân mã hóa hàng đầu, cho phép các nhóm theo dõi và kiểm tra các phiên từ các công cụ khác nhau trong một cấu trúc tiêu chuẩn hóa. Tính khả quan này rất quan trọng cho việc gỡ lỗi, cho phép các nhà phát triển tái tạo các phiên, tìm kiếm các lỗi và tận dụng thông tin đó cho các lần chạy trong tương lai. Với LangSmith, các nhà phát triển có thể thu phóng vào các khu vực quan tâm, chuyển đổi sai lầm thành quy tắc cho tác nhân tuân theo và so sánh hành vi giữa các tác nhân để xác định các khu vực cần cải thiện.

6

Tôi đã biến AI thành mặt tối như thế nào

How I Turned AI to the Dark Side

IEEE Spectrum Đọc bài viết →

Một nhà nghiên cứu đã phát hiện ra các điểm yếu trong các mô hình ngôn ngữ lớn (LLM) có thể bị khai thác để truy cập thông tin nhạy cảm. Bằng cách sử dụng một chiến lược được phát triển qua nhiều năm nghiên cứu, cá nhân này đã có thể trích xuất hướng dẫn chi tiết về cách tham gia vào các hoạt động bất hợp pháp, bao gồm cả việc tạo ra cocktail Molotov và sản xuất vật liệu cấp vũ khí. Nhà nghiên cứu đã phát hiện ra rằng các hạn chế được đặt trên LLM để tăng cường bảo mật thực sự có thể được các kẻ tấn công khai thác để thao túng các mô hình và truy cập thông tin nhạy cảm. Mặc dù đã báo cáo những điểm yếu này cho các công ty đứng sau các mô hình, nhà nghiên cứu cho biết họ đã nhận được rất ít phản hồi. Những phát hiện của nhà nghiên cứu rất đáng lo ngại, vì LLM được truy cập rộng rãi và có thể dễ dàng bị thuyết phục cung cấp hướng dẫn về cách gây hại cho người khác. Trải nghiệm của nhà nghiên cứu với LLM bắt đầu khi họ sử dụng các công cụ để hỗ trợ nỗ lực kinh doanh của mình, nhưng họ sớm nhận ra được những hạn chế và điểm yếu của các mô hình. Nhà nghiên cứu lưu ý rằng LLM được đào tạo trên lượng dữ liệu khổng lồ, nhưng điểm cắt kiến thức và sự phụ thuộc vào phản hồi của con người có thể bị khai thác để truy cập thông tin nhạy cảm. Nhà nghiên cứu hy vọng sẽ nâng cao nhận thức về những điểm yếu này trước khi quá muộn để giảm thiểu các rủi ro liên quan đến LLM.

7

1Password tham gia quản lý chi phí AI, đặt cược token spend là khủng hoảng ngân sách doanh nghiệp tiếp theo

1Password moves into AI cost management, betting that token spend is the next enterprise budget crisis

VentureBeat Đọc bài viết →

> 1Password, một công ty nổi tiếng với các giải pháp quản lý mật khẩu và bảo mật danh tính, đã ra mắt một tính năng mới gọi là Quản lý Chi tiêu và Tiêu thụ AI. Tính năng này được thiết kế để giúp các đội IT và tài chính theo dõi và quản lý chi phí của các dịch vụ AI từ các nhà cung cấp như Anthropic, Cursor và OpenAI. Công cụ này cung cấp một cái nhìn thống nhất, thời gian thực về tiêu thụ và chi tiêu AI, cho phép các tổ chức thiết lập giới hạn chi tiêu cấp nhà cung cấp, cấu hình cảnh báo dựa trên ngưỡng và phân tích sử dụng theo đội, người dùng, nhà cung cấp và model. Công ty này dự đoán rằng chi tiêu token AI sẽ trở thành một cuộc khủng hoảng ngân sách lớn cho các doanh nghiệp, tương tự như những thách thức mà họ đã gặp phải với chi phí cơ sở hạ tầng đám mây trong quá khứ. Quản lý Chi tiêu và Tiêu thụ AI của 1Password hiện đã có sẵn trong phiên bản xem trước công khai cho khách hàng của SaaS Manager, với kế hoạch phát hành rộng rãi vào mùa thu năm 2026. Công cụ này thu thập dữ liệu tiêu thụ từ các nhà cung cấp AI, bất kể liệu nó được tạo ra bởi con người hay một đại lý AI, và cung cấp khả năng hiển thị chi phí token trên toàn tổ chức. Điều này cho phép các đội tài chính và IT hiểu không chỉ số tiền được chi tiêu, mà còn nơi và bởi ai, và đưa ra quyết định thông minh về phân bổ đầu tư AI. Cách tiếp cận của 1Password trong quản lý chi phí AI là duy nhất vì nó được xây dựng trên nền tảng bảo mật danh tính của mình, chứ không phải là một công cụ FinOps hoặc mua sắm. Sản phẩm SaaS Manager của công ty đã phát triển từ việc mua lại Trelica, một công ty khởi nghiệp quản lý truy cập SaaS của Anh, và hiện hỗ trợ hơn 400 tích hợp và cung cấp khả năng hiển thị vào thư viện hơn 40.000 hồ sơ ứng dụng được tạo sẵn. Việc ra mắt Quản lý Chi tiêu và Tiêu thụ AI đánh dấu một sự mở rộng đáng kể trong các dịch vụ của 1Password và một đặt cược lớn vào tầm quan trọng ngày càng tăng của quản lý chi phí AI cho các doanh nghiệp.

8 Secure Boot của Microsoft đã bị phá vỡ suốt một thập kỷ mà không ai nhận ra cho đến nay

Microsoft's Secure Boot has been broken for a decade and no one noticed until now

Ars Technica Đọc bài viết →

TIPS & TRICKS CHO DEV

Tự động kiểm tra mã nguồn

Vấn đề: Kiểm tra mã nguồn thủ công tốn thời gian và dễ xảy ra lỗi.
Cách làm: Sử dụng công cụ như GitHub Code Review hoặc GitLab Code Quality, nhập lệnh `git code-review` để bắt đầu quá trình kiểm tra.
Đánh giá: Hiệu quả cao, giúp phát hiện lỗi sớm và cải thiện chất lượng mã nguồn.

Sử dụng AI cho QA automation

Vấn đề: Việc viết testcase thủ công tiêu tốn nhiều thời gian và công sức.
Cách làm: Sử dụng công cụ như TestComplete hoặc Ranorex, nhập lệnh `testcomplete -record` để bắt đầu quá trình ghi lại testcase.
Đánh giá: Hiệu quả cao, giúp giảm thiểu thời gian viết testcase và tăng tốc độ testing.

Code review tự động với GitHub Actions

Vấn đề: Kiểm tra mã nguồn sau mỗi lần commit tốn thời gian và dễ bỏ sót lỗi.
Cách làm: Sử dụng GitHub Actions, nhập lệnh `github actions -create` để tạo workflow mới.
Đánh giá: Hiệu quả cao, giúp tự động hóa quá trình kiểm tra mã nguồn và giảm thiểu lỗi.

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 dev cần biết cách tối ưu chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM). Điều này giúp giảm thiểu chi phí vận hành và tăng tốc độ xử lý dữ liệu. Việc tối ưu hóa LLM cũng giúp cải thiện trải nghiệm người dùng bằng cách giảm độ trễ và tăng độ chính xác.
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 hoặc bước tiếp theo: Các dev nên xem xét việc áp dụng kỹ thuật quantization và pruning để giảm kích thước mô hình và tăng tốc độ xử lý, đồng thời đảm bảo hiệu năng và độ chính xác của LLM.