

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Bàn tay ta làm nên tất cả, có sức người sỏi đá cũng thành cơm.”

— Ca dao Việt Nam

Lao động chân chính là nguồn gốc của mọi thành quả — bằng sức mình, con người có thể vượt qua mọi nghịch cảnh.

TIN TỨC NỔI BẬT

1

11 lựa chọn thay thế GitHub Copilot mà developer đã chuyển sang dùng năm 2026

11 GitHub Copilot Alternatives Developers Switched To in 2026

Security Boulevard

[Đọc bài viết →](#)

Năm 2026, các developer đã tìm kiếm những lựa chọn thay thế cho GitHub Copilot, một AI coding assistant phổ biến. Kết quả là, một số lựa chọn thay thế đã trở nên phổ biến trong cộng đồng developer. Theo các xu hướng gần đây, các developer đã chuyển sang dùng 11 lựa chọn thay thế GitHub Copilot. Các lựa chọn này cung cấp nhiều tính năng và chức năng khác nhau, đáp ứng các nhu cầu và sở thích đa dạng. Một số lựa chọn thay thế đáng chú ý bao gồm TabNine, cung cấp các gợi ý và hoàn thành code được hỗ trợ bởi AI, và Kite, một coding assistant cung cấp tính năng hoàn thành và gợi ý code theo thời gian thực. Các lựa chọn khác, như CodeWithMe, cung cấp các tính năng collaborative coding, trong khi một số khác, như DeepCode, tập trung vào code analysis và bảo mật. Ngoài ra, các lựa chọn thay thế như Code Scanner của chính GitHub và CodeFactor, cung cấp các công cụ code review và analysis. Các lựa chọn thay thế đáng chú ý khác bao gồm Codiga, cung cấp tính năng hoàn thành và gợi ý code được hỗ trợ bởi AI, và CodePro, cung cấp tính năng hoàn thành và gợi ý code theo thời gian thực.

2

Tình hình các Agent Framework: Chọn Runtime phù hợp để triển khai AI trong doanh nghiệp

State of Agent Frameworks: Choosing the Right Runtime for Enterprise AI Execution

Bài viết "State of Agent Frameworks: Choosing the Right Runtime for Enterprise AI Execution" cung cấp cái nhìn tổng quan về tình hình hiện tại của các agent framework trong bối cảnh triển khai AI trong doanh nghiệp. Agent framework là các nền tảng phần mềm cho phép phát triển và triển khai các intelligent agent, vốn là các thực thể phần mềm tự động có khả năng tương tác với môi trường và đưa ra quyết định. Bài viết nhấn mạnh tầm quan trọng của việc chọn đúng runtime để triển khai AI trong doanh nghiệp, vì nó có thể ảnh hưởng đáng kể đến performance, scalability và reliability của các ứng dụng AI. Bài viết thảo luận về nhiều agent framework khác nhau, bao gồm Rasa, Microsoft Bot Framework và Google Dialogflow, cùng với những điểm mạnh và điểm yếu tương ứng của chúng. Bài viết cũng đề cập đến các yếu tố chính cần cân nhắc khi lựa chọn một agent framework, như ease of use, khả năng customization và integration với các hệ thống hiện có. Nó nhấn mạnh sự cần thiết của một framework có thể xử lý các cuộc hội thoại phức tạp, cung cấp bảo mật mạnh mẽ và hỗ trợ tương tác đa kênh (multi-channel). Bằng cách chọn đúng agent framework, các doanh nghiệp có thể xây dựng các ứng dụng AI hiệu quả hơn, mang lại giá trị kinh doanh và cải thiện trải nghiệm khách hàng.

3

Xây dựng workflow được hỗ trợ bởi agent dùng Cortex AI và Managed MCP Servers

Build Agent-Powered Workflows Using Cortex AI and Managed MCP Servers

Snowflake

[Đọc bài viết →](#)

Snowflake đã giới thiệu một khả năng mới cho phép người dùng xây dựng các workflow được hỗ trợ bởi agent sử dụng Cortex AI và Managed MCP servers. Sự tích hợp này cho phép các developer tận dụng sức mạnh của Cortex AI để tự động hóa và tối ưu hóa các workflow phức tạp, đồng thời hưởng lợi từ scalability và reliability của Managed MCP servers của Snowflake. Với khả năng mới này, người dùng có thể tạo các custom workflow để tự động hóa các tác vụ, xử lý dữ liệu và analytics, tất cả đều trong nền tảng Snowflake. Sự tích hợp giữa Cortex AI và Managed MCP servers cung cấp một hạ tầng mạnh mẽ và scalable để xây dựng và triển khai các workflow này, đảm bảo performance cao và reliability. Khả năng này được thiết kế để giúp các tổ chức tinh gọn các data workflow của họ, cải thiện hiệu quả và giảm chi phí. Bằng cách tự động hóa các tác vụ lặp đi lặp lại và tối ưu hóa việc xử lý dữ liệu, người dùng có thể tập trung vào các tác vụ có giá trị

cao hơn và thúc đẩy business insights. Sự tích hợp giữa Cortex AI và Managed MCP servers là một bước tiến đáng kể trong sứ mệnh của Snowflake nhằm cung cấp một data platform linh hoạt và scalable cho các doanh nghiệp hiện đại.

4

Đánh giá OneOdio Studio Max 2: Tai nghe độ trễ cực thấp dành cho dân công nghệ

OneOdio Studio Max 2 review: Ultra-low latency headphones for tech enthusiasts

BD Tech Talks

[Đọc bài viết →](#)

OneOdio Studio Max 2 là tai nghe wireless được thiết kế dành cho tech enthusiast, developer và gamer, những người yêu cầu độ trễ thấp (low latency) và khả năng kết nối linh hoạt. Những chiếc tai nghe closed-back, over-ear này kết hợp nhiều chế độ wireless và wired, loại bỏ độ trễ âm thanh thường thấy ở các tai nghe Bluetooth tiêu chuẩn. Chúng có một transmitter 2.4GHz cho các phiên chơi game tốc độ cao, chỉnh sửa video và coding. Tai nghe có thiết kế thoải mái với earpad bằng memory foam và headband kim loại có đánh số để điều chỉnh kích thước chính xác. Chúng hỗ trợ bốn phương thức kết nối, bao gồm Bluetooth 6.0, một M2 USB transmitter cho âm thanh không độ trễ (

5

Allstate cáo buộc Broadcom kiểm toán vì đã bỏ VMware, CA

Allstate accuses Broadcom of auditing it because it quit VMware, CA

Ars Technica

[Đọc bài viết →](#)

Công ty Bảo hiểm Allstate đã cáo buộc Broadcom đã thực hiện các cuộc kiểm toán đối với họ như một hình thức trả thù vì quyết định chấm dứt hợp đồng của công ty bảo hiểm với VMware và CA Technologies. Những cáo buộc này được đưa ra trong một vụ kiện được VMware đệ trình chống lại Allstate vào tháng 12 năm 2025. Allstate tuyên bố rằng Broadcom đã khởi xướng các cuộc kiểm toán sau khi họ quyết định không gia hạn hợp đồng với các công ty này, với lý do họ đã tuân thủ đáng kể các yêu cầu kiểm toán và báo cáo được quy định trong hợp đồng với VMware. Việc Broadcom tiếp quản VMware được cho là đã thúc đẩy các công ty khác, bao gồm T-Mobile, Tesco và Western Union, chuyển khỏi công nghệ ảo hóa. Tranh chấp giữa Allstate và Broadcom vẫn đang diễn ra, với cả hai bên có thời hạn đến ngày 17 tháng 5 năm 2027 để nộp các đơn kiến nghị quyết. Vụ

việc này làm nổi bật tính cách sẵn sàng kiện tụng của Broadcom và sự sẵn sàng chiến đấu với các khách hàng VMware không hài lòng.

6

Cá nhân chịu trách nhiệm trực tiếp (DRI)

Directly Responsible Individuals (DRI)

Simon Willison

[Đọc bài viết →](#)

Khái niệm về Các Cá Nhân Trực Tiếp Chịu Trách Nhiệm (DRI) đã được khám phá trong bối cảnh các tổ chức của con người và trí tuệ nhân tạo. Theo sổ tay của GitLab, thuật ngữ này bắt nguồn từ Apple, nơi nó đề cập đến người cuối cùng chịu trách nhiệm về thành công hoặc thất bại của một dự án, sáng kiến hoặc hoạt động cụ thể. Trách nhiệm này được coi là duy nhất của con người, vì các cá nhân có thể chịu trách nhiệm về hành động của mình, trong khi máy móc không thể. Ý tưởng này được củng cố bởi một trang trình bày đào tạo năm 1979 của IBM, tuyên bố rằng máy tính không nên đưa ra quyết định quản lý do khả năng không thể bị ràng buộc trách nhiệm. Điều này đặt ra câu hỏi về vai trò của các tác nhân được hỗ trợ bởi LLM trong các tổ chức của con người và liệu chúng có nên được coi là DRI cho các dự án.

7

OpenWiki Brains: Bộ nhớ chủ động cho AI Agent

OpenWiki Brains: Proactive Memory for AI Agents

LangChain Blog

[Đọc bài viết →](#)

OpenWiki Brains là một framework mới cho phép các tác nhân AI có bộ nhớ chủ động. Không giống như các giải pháp bộ nhớ hiện có dựa trên cập nhật rõ ràng hoặc các mẫu được phát hiện thông qua sử dụng, OpenWiki Brains cho phép các tác nhân xây dựng bộ nhớ bằng cách lấy bối cảnh liên quan từ các công cụ, kênh và tệp tin kết nối. Framework này mở rộng khái niệm OpenWiki, một CLI cho tài liệu mã nguồn, và hiện có thể tạo ra một bộ não đa năng cho các tác nhân. OpenWiki Brains kết nối với các nguồn khác nhau như Gmail, Notion và tìm kiếm web, và chuyển đổi thông tin thành một wiki cục bộ mà các tác nhân có thể sử dụng làm bộ nhớ. Mục tiêu là cung cấp cho các tác nhân quyền truy cập vào bối cảnh mới mà không cần theo dõi hoặc bảo trì thủ công. Điều này đặc biệt hữu ích cho các tác nhân làm việc trên nhiều nhiệm vụ, chẳng hạn như nghiên cứu, lập kế hoạch và công việc khách hàng, nơi bối cảnh hữu ích bị phân tán trên nhiều nơi. OpenWiki Brains nhằm mục đích cung cấp cho các tác nhân một nơi

bền vững để tìm kiếm, biến các nguồn kết nối thành một wiki có cấu trúc có thể được làm mới theo thời gian.

8

Sau Orthogonality: Agency đạo đức đức hạnh và AI Alignment

After Orthogonality: Virtue-Ethical Agency and AI Alignment

The Gradient

[Đọc bài viết →](#)

Bài viết "Sau Tính Chính Orthogonal: Cơ quan Đạo đức - Đức hạnh và Sự Liên kết AI" thách thức cách tiếp cận truyền thống đối với sự liên kết trí tuệ nhân tạo (AI) bằng cách lập luận rằng con người hợp lý và AI không nên có mục tiêu. Thay vào đó, hành động của con người được hướng dẫn bởi các thực hành, là mạng lưới các hành động, khuynh hướng, tiêu chí đánh giá và tài nguyên cấu trúc và thúc đẩy chúng. Tác giả đề xuất rằng sự suy xét của các tác nhân AI nên chia sẻ logic dựa trên thực hành này để thực sự hỗ trợ sự hoạt động của con người. Cách tiếp cận này rất quan trọng để liên kết AI với các thuộc tính an toàn như minh bạch, hữu ích và vô hại. Bài viết cũng khám phá khái niệm eudaimonia, hay sự thịnh vượng con người tích cực, hợp lý, và lập luận rằng nó chỉ ra một cấu trúc suy xét khác với tính hợp lý kết quả tiêu chuẩn. Tác giả cho rằng tính hợp lý eudaimonic này là một khuôn khổ hữu ích cho AI được liên kết với con người, vì nó có thể cung cấp lợi thế ổn định và an toàn. Bài viết kết luận rằng nếu tính hợp lý eudaimonic là tự nhiên và hiệu quả, nó có thể là một cách tiếp cận khả thi để liên kết AI với sự thịnh vượng của con người, giải quyết các vấn đề an toàn AI cổ điển và nghịch lý.

TIPS & TRICKS CHO DEV

Chain-of-thought Prompting

Vấn đề: Model AI không hiểu được context của câu hỏi phức tạp.

Cách làm: Sử dụng chain-of-thought prompting, ví dụ "Break down the solution into steps: ...".

Đánh giá: Hiệu quả khi giải quyết bài toán phức tạp, không nên dùng với câu hỏi đơn giản.

Few-shot Learning

Vấn đề: Model AI cần nhiều dữ liệu để học tập.

Cách làm: Cung cấp vài ví dụ liên quan, ví dụ "Given the following examples: ...".

Đánh giá: Hiệu quả khi có ít dữ liệu, không nên dùng khi có nhiều dữ liệu.

System Prompt Design

Vấn đề: Model AI không hiểu được yêu cầu cụ thể.

Cách làm: Định nghĩa rõ hệ thống và yêu cầu, ví dụ "You are a chatbot, answer: ...".

Đánh giá: Hiệu quả khi cần trả lời chính xác, nên dùng với yêu cầu cụ thể.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Để phát triển ứng dụng AI hiệu quả, 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 cải thiện trải nghiệm người dùng. Việc tối ưu hóa cũng cho phép tích hợp LLM vào các ứng dụng di động và edge devices.

3. Ví dụ, sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation of Weights) để giảm kích thước mô hình và tăng tốc độ xử lý.

4. Tip hoặc bước tiếp theo: Hãy bắt đầu bằng cách đánh giá nhu cầu và nguồn lực của dự án, sau đó chọn kỹ thuật tối ưu phù hợp để đạt được hiệu suất và chi phí mong muốn. Sử dụng các công cụ như Hugging Face Transformers để triển khai và tối ưu hóa LLM.

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

Nguồn: Google News · Groq AI