

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Gieo gió gặt bão.”

— Tục ngữ Việt Nam

Mọi hành động đều có hệ quả — làm điều tốt sẽ nhận lại điều tốt, làm điều xấu sẽ gánh chịu hậu quả.

TIN TỨC NỔI BẬT

1

Developer thử nghiệm cuộc sống sau khi GitHub Copilot thay đổi chính sách tính phí

Developers are testing life after GitHub Copilot changes its billing

Startup Fortune

[Đọc bài viết →](#)

Các nhà phát triển hiện đang thử nghiệm cuộc sống sau khi GitHub Copilot thay đổi cách tính phí. GitHub Copilot là một công cụ mã hóa được hỗ trợ bởi AI, giúp các nhà phát triển với đề xuất mã và hoàn thiện mã. Trong một động thái gần đây, GitHub đã thay đổi mô hình tính phí cho Copilot, chuyển từ phí hàng tháng cố định sang hệ thống giá dựa trên sử dụng. Thay đổi này đã khiến một số nhà phát triển phải đánh giá lại việc sử dụng công cụ của họ. Kết quả là, các nhà phát triển đang khám phá các công cụ và dịch vụ mã hóa thay thế để đáp ứng nhu cầu của họ. Một số người chọn các giải pháp thay thế miễn phí hoặc thấp chi phí, trong khi những người khác đang tìm kiếm các lựa chọn tiết kiệm hơn cung cấp chức năng tương tự như Copilot. Thay đổi cách tính phí cũng đã dẫn đến sự tập trung lại vào giá trị của Copilot và các đối thủ cạnh tranh của nó trên thị trường mã hóa được hỗ trợ bởi AI. Giai đoạn thử nghiệm này cho phép các nhà phát triển đánh giá tác động của thay đổi cách tính phí đối với quy trình làm việc và năng suất của họ. Hiện vẫn chưa rõ sự thay đổi này sẽ ảnh hưởng như thế nào đến việc áp dụng và sử dụng GitHub Copilot và các công cụ mã hóa được hỗ trợ bởi AI khác.

2

Hiện trạng Agent Frameworks: Chọn Runtime phù hợp cho triển khai AI trong doanh nghiệp

Bài viết "Trạng thái của các Framework Trình điều khiển: Chọn Thời gian chạy Phù hợp cho Thực thi AI Doanh nghiệp" khám phá cảnh quan hiện tại của các framework trình điều khiển trong thực thi AI doanh nghiệp. Các framework trình điều khiển là hệ thống phần mềm cho phép tích hợp các model trí tuệ nhân tạo (AI) và học máy (ML) vào các ứng dụng kinh doanh. Bài viết nhấn mạnh tầm quan trọng của việc chọn thời gian chạy phù hợp cho các framework này, vì nó có thể ảnh hưởng đáng kể đến hiệu suất và khả năng mở rộng của các hệ thống được ☐ động bởi AI. Bài viết thảo luận về các framework trình điều khiển khác nhau, bao gồm Rasa, Dialogflow và Microsoft Bot Framework, cũng như các điểm mạnh và điểm yếu tương ứng của chúng. Nó cũng đề cập đến các thách thức khi chọn framework phù hợp, bao gồm tính tương thích, khả năng mở rộng và tích hợp với các hệ thống hiện có. Tác giả nhấn mạnh nhu cầu về một framework có thể xử lý các cuộc trò chuyện phức tạp, hỗ trợ nhiều ngôn ngữ và cung cấp tích hợp liền mạch với các ứng dụng doanh nghiệp. Cuối cùng, bài viết nhằm cung cấp hướng dẫn cho các nhà phát triển doanh nghiệp và kỹ sư AI trong việc chọn framework trình điều khiển phù hợp nhất cho nhu cầu thực thi AI của họ. Bằng cách hiểu trạng thái hiện tại của các framework trình điều khiển và khả năng của chúng, các tổ chức có thể đưa ra quyết định thông minh và xây dựng các hệ thống được ☐ động bởi AI hiệu quả hơn.

3

Xây dựng Workflow dùng Agent với 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 tích hợp mới cho phép người dùng xây dựng các quy trình làm việc được hỗ trợ bởi các tác nhân sử dụng Cortex AI và máy chủ Managed MCP. Tích hợp này cho phép các nhà phát triển tận dụng sức mạnh của Cortex AI, nền tảng AI và học máy của Snowflake, để tự động hóa các nhiệm vụ và quy trình làm việc phức tạp. Với tích hợp này, người dùng có thể tạo các quy trình làm việc tùy chỉnh sử dụng các khả năng của Cortex AI, chẳng hạn như xử lý dữ liệu và đào tạo model. Các quy trình làm việc có thể được kích hoạt bởi các sự kiện, lịch trình hoặc tương tác của người dùng và có thể được quản lý thông qua máy chủ Managed MCP của Snowflake. Các quy trình làm việc được hỗ trợ bởi tác nhân có thể được sử dụng

để tự động hóa một loạt các nhiệm vụ, từ việc nhập và xử lý dữ liệu đến việc triển khai và giám sát model. Tích hợp này nhằm mục đích đơn giản hóa quá trình xây dựng và quản lý quy trình làm việc, giúp các nhà phát triển tập trung vào các nhiệm vụ phức tạp hơn. Bằng cách kết hợp Cortex AI với máy chủ Managed MCP, Snowflake cung cấp một nền tảng có thể mở rộng và bảo mật để xây dựng và triển khai quy trình làm việc, cho phép người dùng tự động hóa các nhiệm vụ lặp đi lặp lại và cải thiện năng suất tổng thể.

4

ASPIRE framework của Nvidia tăng tốc lập trình robot bằng AI tự cải thiện như thế nào

How Nvidia's ASPIRE framework accelerates robot programming with self-improving AI

BD Tech Talks

[Đọc bài viết →](#)

Nvidia đã phát triển khuôn khổ ASPIRE hợp tác với các nhà nghiên cứu từ nhiều trường đại học, nhằm cách mạng hóa lập trình robot truyền thống. Khuôn khổ này sử dụng phương pháp AI tự cải tiến để viết, thực thi và tinh chỉnh chương trình điều khiển robot một cách tự động. ASPIRE chẩn đoán lỗi của chính nó và tạo ra một thư viện kỹ năng có thể tái sử dụng, giảm đáng kể nỗ lực lập trình, token và gỡ lỗi cần thiết trên robot vật lý. Kỹ thuật này phù hợp với xu hướng ngành, nơi các tác nhân AI cải tiến chính khung xương của mình, và rất liên quan đến các ứng dụng AI trong thế giới thực. Bằng cách tích lũy kiến thức có thể chuyển giao, ASPIRE giúp tăng tốc quá trình thích nghi và triển khai hệ thống robot trong môi trường phức tạp. Khuôn khổ này hoạt động như một kỹ sư robot cao cấp, liên tục học hỏi từ sai lầm và tinh chỉnh kỹ năng của mình. Phương pháp này trái ngược với các hệ thống hiện tại, không tích lũy kinh nghiệm và gặp khó khăn trong việc xác định chiến lược sửa chữa hiệu quả. Phương pháp học liên tục của ASPIRE có tiềm năng mang lại những cải tiến đáng kể trong lập trình và triển khai robot.

5

MCP trên Code Mode

MCP on Code Mode

Changelog

[Đọc bài viết →](#)

Trong tuần này, Matt Carey từ Cloudflare thảo luận về Code Mode và mối quan hệ của nó với MCP (Model-Code-Protocol). Carey tiết lộ rằng hầu hết mọi người đã nghĩ về MCP một cách không chính xác. Ông

giải thích cách Code Mode phía server cho phép một máy chủ MCP duy nhất $\square$ hơn 2.500 điểm cuối API của Cloudflare bằng cách sử dụng khoảng 1.000 token ngữ cảnh. Ngoài ra, Carey cũng nói về trình tải Worker động mà chạy an toàn mã được viết bởi model trong một V8 isolate. Cuộc trò chuyện cũng đề cập đến quy trình làm việc của Carey với Claude, vai trò của bộ nhớ trong tương lai của các agent và việc ông sử dụng một công cụ bao git gọi là Zaggy để ngăn chặn force-push tới các kho lưu trữ của mình.

6

Thỏa thuận bản quyền 1,5 tỷ USD mang tính bước ngoặt của Anthropic được phê duyệt

Anthropic's landmark \$1.5B copyright settlement is approved

TechCrunch AI

[Đọc bài viết →](#)

Một thẩm phán liên bang đã cấp phê duyệt cuối cùng cho việc giải quyết 1,5 tỷ đô la của Anthropic trong một vụ kiện tập thể về bản quyền. Vụ kiện, liên quan đến các tác giả và nhà xuất bản sách, cáo buộc Anthropic đã tải xuống và lưu trữ bất hợp pháp hàng triệu cuốn sách có bản quyền. as phần của việc giải quyết, Anthropic sẽ trả 3.000 đô la cho mỗi tác phẩm trên ước tính 500.000 tác phẩm. Tuy nhiên, việc giải quyết này không giải quyết vấn đề cốt lõi về việc liệu việc đào tạo một AI model trên văn bản có bản quyền có được coi là sử dụng hợp lý hay không. Một phán quyết trước đó của Thẩm phán William Alsup cho thấy phương pháp đào tạo AI model của Anthropic là sử dụng hợp lý, nhưng phương pháp của công ty trong việc có được các cuốn sách, bao gồm tải xuống từ các trang web hải tặc, là bất hợp pháp. Phán quyết này đã được coi là một bước ngoặt cho ngành công nghiệp AI, nhưng tác động của nó bị hạn chế vì đây là quyết định của một tòa án khu vực và không phải là tiền lệ ràng buộc. Các công ty khác, như Google, Meta, Midjourney và OpenAI, vẫn phải đối mặt với các vụ kiện bản quyền về việc sử dụng các tác phẩm có bản quyền để đào tạo các AI model.

7

Galaxy Card là câu trả lời của Samsung dành cho Apple Card

The Galaxy Card Is Samsung's Answer to the Apple Card

Wired

[Đọc bài viết →](#)

Samsung đã công bố Galaxy Card, một thẻ tín dụng nhằm cạnh tranh với thẻ Apple Card của Apple. Được phát hành bởi Barclays trên mạng

Visa, Galaxy Card cung cấp phần thưởng cho các mua hàng được thực hiện thông qua Samsung, bao gồm 5% tiền mặt hoàn lại cho mua hàng trong cửa hàng hoặc trực tuyến, 3% cho mua hàng sử dụng Samsung Wallet và 2% cho đăng ký dịch vụ phát trực tuyến. Thẻ này không có phí hàng năm và không có phí giao dịch nước ngoài, với lãi suất hàng năm (APR) thay đổi theo từng thành viên thẻ. Để nhận được phần thưởng, thành viên thẻ phải sử dụng thẻ vật lý hoặc thẻ ảo được cấp cho tài khoản Samsung Wallet của họ. Galaxy Card cũng cung cấp giảm giá 20% cho chương trình thành viên VIP Advantage của Samsung và 200 đô la tiền mặt hoàn lại sau khi chi tiêu 2.000 đô la trong 90 ngày đầu tiên. Thẻ này có sẵn cho bất kỳ ai, không chỉ chủ sở hữu thiết bị Samsung, và có thể được quản lý thông qua cổng trực tuyến của Barclays. Các chuyên gia trong ngành tin rằng Galaxy Card là một bước đi chiến lược của Samsung để tăng cường sự hiện diện và lòng trung thành của thương hiệu, chứ không phải là một sản phẩm cách mạng.

8

Các AI model của Trung Quốc khiến thế giới AI của Trump tự đấu đá nội bộ

China's AI models have Trump's AI world at war with itself

MIT Tech Review

[Đọc bài viết →](#)

Các mô hình AI của Trung Quốc đã gây ra một cuộc tranh luận nóng tại Mỹ, đặc biệt là среди các cố vấn của cựu Tổng thống Donald Trump. Việc phát hành Kimi, một mô hình AI miễn phí và mã nguồn mở từ công ty Moonshot của Trung Quốc, đã được coi là một mối đe dọa đối với các công ty Mỹ như Anthropic và OpenAI. Những công ty Mỹ này thu phí để truy cập vào các mô hình AI của họ, trong khi Kimi và các mô hình tương tự của Trung Quốc là miễn phí. Điều này đã dẫn đến một sự chia rẽ giữa các cố vấn của Trump, với một số người cho rằng chính phủ nên can thiệp để bảo vệ các công ty Mỹ, trong khi những người khác tin rằng AI mở hơn là tốt hơn. Cuộc tranh luận đã trở nên cá nhân, với các cố vấn hiện tại và trước đây trao đổi những lời xúc phạm. David Sacks, cố vấn AI và tiền điện tử trước đây của Trump, đã gọi các mô hình của Anthropic là "đã bị cắt bỏ một phần não" và "thức tỉnh", trong khi Emil Michael, một quan chức cấp cao của Lầu Năm Góc, đã gọi người đứng đầu tương lai chiến lược của OpenAI là "người gốc tối thượng". Sự không đồng ý này cũng đã gây ra những lo ngại về tác động kinh tế của các mô hình AI của Trung Quốc, điều này có thể dẫn đến sự suy giảm hứng thú với các công ty Mỹ và mất tăng trưởng kinh tế. Tình hình đã trở nên phức tạp hơn bởi lệnh cấm xây

dựng các trung tâm dữ liệu mới tại New York và sự mất lòng tin ngày càng tăng đối với các công ty AI. Kết quả là, Tổng thống Trump đang phải đối mặt với một quyết định khó khăn về cách giải quyết vấn đề này. Một số người đã đề xuất rằng ông có thể sử dụng quyền lực mềm để ngăn chặn các công ty Mỹ sử dụng các mô hình AI của Trung Quốc, trong khi những người khác tin rằng chính phủ nên đóng vai trò tích cực hơn trong việc điều chỉnh sự phát triển của AI.

TIPS & TRICKS CHO DEV

Sử dụng MCP cho truy cập file

Vấn đề: Cần truy cập file hệ thống từ trong ứng dụng AI.

Cách làm: Sử dụng Model Context Protocol (MCP) với Claude Desktop, kết hợp lệnh `file://` để truy cập file. Ví dụ, `file://path/to/file.txt`.

Đánh giá: Hiệu quả cho ứng dụng cần truy cập file cục bộ.

Tích hợp MCP với Git

Vấn đề: Cần tích hợp AI với hệ thống Git.

Cách làm: Sử dụng MCP với Claude Code, lệnh `git://` để kết nối với Git. Ví dụ, `git://github.com/repo`.

Đánh giá: Tăng hiệu suất cho dự án cần tích hợp AI và Git.

MCP cho database và APIs

Vấn đề: Cần truy cập database và APIs từ ứng dụng AI.

Cách làm: Sử dụng MCP với Claude Code, kết hợp lệnh `db://` và `api://` để truy cập database và APIs. Ví dụ, `db://mysql/db` và `api://api.example.com`.

Đánh giá: Hiệu quả cho ứng dụng cần truy cập dữ liệu từ nhiều nguồn.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Để tối ưu hóa chi phí và hiệu năng của các mô hình ngôn ngữ lớn (LLM), các nhà phát triển cần biết cách giảm thiểu tài nguyên tính toán và tối ưu hóa hiệu suất.

Điều này giúp giảm chi phí vận hành và tăng tốc độ phản hồi của ứng dụng. Các kỹ thuật như fine-tuning và LoRA có thể giúp đạt được mục tiêu này.

3. Ví dụ, bằng cách sử dụng kỹ thuật fine-tuning, chúng ta có thể điều chỉnh mô hình LLM để phù hợp với một nhiệm vụ cụ thể, giảm thiểu sự phức tạp và tăng tốc độ xử lý.

4. Tip hoặc bước tiếp theo: Hãy bắt đầu bằng cách đánh giá nhu cầu và giới hạn tài nguyên của ứng dụng, sau đó chọn kỹ thuật tối ưu hóa phù hợp để cải thiện hiệu suất và giảm chi phí.

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

Nguồn: Google News · Groq AI