

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Người ta là hoa đất.”

— Tục ngữ Việt Nam

Con người là tạo vật quý giá nhất — mỗi người đều mang trong mình giá trị và tiềm năng độc đáo.

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 mà không có mô hình định giá ban đầu của GitHub Copilot, sau những thay đổi gần đây về cấu trúc hóa đơn của nó. GitHub Copilot là một trợ lý mã hóa được hỗ trợ bởi AI, cung cấp cho các nhà phát triển đề xuất và hoàn thiện mã. Công cụ này đã là một yếu tố thay đổi cuộc chơi cho nhiều lập trình viên, nhưng mô hình định giá trước đó của nó, tính phí người dùng dựa trên số dòng mã họ viết, đã vấp phải sự chỉ trích. Cấu trúc hóa đơn mới cho GitHub Copilot dựa trên phí đăng ký hàng tháng, với người dùng có tùy chọn trả hàng năm để được giảm giá. Thay đổi này đã gây ra phản ứng lẫn lộn từ các nhà phát triển, với một số chào đón mô hình định giá mới như minh bạch hơn và những người khác bày tỏ lo ngại về chi phí tăng lên. Khi các nhà phát triển thích nghi với mô hình định giá mới, nhiều người đang khám phá các công cụ mã hóa được hỗ trợ bởi AI thay thế, chẳng hạn như Kite và TabNine. Những công cụ này cung cấp chức năng tương tự như GitHub Copilot nhưng với cấu trúc định giá khác, cho phép các nhà phát triển tìm kiếm lựa chọn tốt nhất cho nhu cầu và ngân sách của họ.

2

Tình trạng các Agent Framework: Chọn Runtime phù hợp cho việc triển khai AI trong doanh nghiệp

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

Medium

[Đọc bài viết →](#)

Bài viết "Trạng thái của các Framework Trợ lý: Chọn Runtime Đúng cho Thực thi AI Doanh nghiệp" khám phá trạng thái hiện tại của các framework trợ lý trong bối cảnh thực thi AI doanh nghiệp. Các framework trợ lý là các nền tảng phần mềm cho phép phát triển và triển khai các model trí tuệ nhân tạo (AI). Bài viết nhấn mạnh tầm quan trọng của việc chọn runtime đúng cho thực thi AI doanh nghiệp, vì nó có thể ảnh hưởng đáng kể đến hiệu suất, khả năng mở rộng và độ tin cậy của các ứng dụng AI. Bài viết thảo luận về các framework trợ lý khác nhau, bao gồm Rasa, Dialogflow và Botpress, cũng như các điểm mạnh và điểm yếu tương ứng của chúng. Nó cũng đề cập đến tầm quan trọng của việc xem xét các yếu tố như sự dễ sử dụng, tùy chọn tùy chỉnh và tích hợp với các hệ thống hiện có khi chọn một framework trợ lý. Hơn nữa, bài viết nhấn mạnh nhu cầu về một runtime linh hoạt và có khả năng mở rộng có thể đáp ứng nhu cầu phát triển của các ứng dụng AI doanh nghiệp. Bằng cách cung cấp cái nhìn tổng quan về trạng thái hiện tại của các framework trợ lý, bài viết nhằm giúp các nhà phát triển và tổ chức đưa ra quyết định sáng suốt khi chọn một runtime cho các dự án AI của họ.

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 khả năng 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 tạo và tự động hóa các quy trình làm việc phức tạp, đơn giản hóa các nhiệm vụ và nâng cao năng suất. Với Cortex AI, người dùng có thể tận dụng học máy và trí tuệ nhân tạo để phân tích dữ liệu và đưa ra quyết định thông minh. Bằng cách kết hợp Cortex AI với máy chủ Managed MCP, các nhà phát triển có thể tạo các quy trình làm việc tùy chỉnh tự động hóa các nhiệm vụ lặp đi lặp lại, giải phóng tài nguyên cho các sáng kiến chiến lược hơn. Tích hợp này cung cấp một số lợi ích, bao gồm hiệu quả cải thiện, khả năng mở rộng tăng cường và tính linh hoạt tăng cao. Người dùng hiện có thể tạo và triển khai các quy trình làm việc thích ứng với nhu cầu kinh doanh thay đổi, đảm bảo rằng dữ liệu và ứng dụng của họ luôn cập nhật và bảo mật. Các quy trình làm việc được hỗ trợ bởi tác nhân của Snowflake được thiết kế để có thể tùy chỉnh cao, cho phép người dùng điều chỉnh quy

trình làm việc của họ cho các yêu cầu kinh doanh cụ thể. Khả năng này dự kiến sẽ nâng cao trải nghiệm người dùng tổng thể và thúc đẩy tăng trưởng kinh doanh bằng cách tự động hóa các nhiệm vụ phức tạp và cải thiện việc ra quyết định dựa trên dữ liệu.

4

Firefox trên WebAssembly

Firefox in WebAssembly

Simon Willison

[Đọc bài viết →](#)

Firefox đã được biên dịch thành công để chạy trong WebAssembly, cho phép toàn bộ trình duyệt chạy trong một trình duyệt khác. Dự án này, do Puter dẫn đầu, đã sử dụng trình duyệt Firefox/Gecko do nó có hỗ trợ mạnh mẽ cho quá trình đơn. Quá trình biên dịch này được thực hiện thông qua việc sử dụng Claude Opus và token Fable, với chi phí ước tính là 25.000 đô la. Để thực hiện điều này, một máy chủ proxy đã được sử dụng để chuyển tất cả lưu lượng truy cập qua giao thức WebSocket, đảm bảo rằng mã chạy trong trình duyệt không thể tạo kết nối mạng tùy ý. Bản demo trình duyệt Firefox đã được biên dịch chạy trong Chrome, với nhóm tuyên bố rằng nó hỗ trợ mã hóa từ đầu đến cuối. Kho lưu trữ của dự án có sẵn để xem, và một dự án tương tự, theogbob/WebkitWasm, cũng đã được khởi xướng để biên dịch WebKit sang WASM.

5

Giám sát kỹ thuật số định hình lại việc thực thi pháp luật thủy sản ở Indonesia

Digital Surveillance Reshapes Fishery Enforcement in Indonesia

IEEE Spectrum

[Đọc bài viết →](#)

Indonesia đang cách mạng hóa cách tiếp cận thực thi pháp luật về đánh bắt cá với sự giúp đỡ của giám sát kỹ thuật số. Quốc gia này đang sử dụng một loạt các công cụ cảm biến từ xa và phân tích để theo dõi hoạt động đánh bắt cá trong vùng biển rộng lớn của mình. Với sự hỗ trợ của theo dõi vệ tinh và hệ thống giám sát tàu thuyền, các cơ quan chức năng Indonesia hiện có thể xác định các vi phạm tiềm năng trong thời gian thực, thậm chí trước khi tàu tuần tra được triển khai. Sự thay đổi này trong quản lý hàng hải làm cho việc thực thi pháp luật trên một khu vực rộng lớn hơn trở nên khả thi, vì đại dương không còn trong suốt với các nhà quản lý. Việc sử dụng các công cụ giám sát kỹ thuật số, chẳng hạn như Global Fishing Watch, đã làm cho

việc theo dõi hàng trăm nghìn tàu thuyền trên toàn thế giới trở nên khả thi, cung cấp một bức tranh gần như thời gian thực về hoạt động đánh bắt cá. Cách tiếp cận đầy tham vọng của Indonesia là một trong những ví dụ đáng chú ý nhất về sự chuyển đổi này, vì quốc gia này quản lý hơn sáu triệu kilômét vuông không gian hàng hải với nguồn lực thực thi hạn chế. Việc tích hợp giám sát kỹ thuật số vào thực thi pháp luật về đánh bắt cá đã làm tăng đáng kể số lượng tàu thuyền được theo dõi, với 9.394 tàu cá Indonesia hiện đang truyền tải tích cực thông qua Hệ thống Giám sát Tàu thuyền Quốc gia.

6

HP bị phạt 1,4 tỷ rupee vì "thỏa thuận độc quyền" hộp mực, toner và PC

HP fined 1.4 billion rupees for "cartelization" of ink cartridges, toner, PCs

Ars Technica

[Đọc bài viết →](#)

Chính phủ Ấn Độ đã phạt HP Ấn Độ và các đối tác của họ với tổng số tiền là 1,4 tỷ rupee (khoảng 14,4 triệu đô la) vì tham gia vào "cartelization" của máy tính, hộp mực và bột mực. Ủy ban Cạnh tranh Ấn Độ (CCI) đã phát hiện ra rằng HP Ấn Độ đã thông đồng với một số đối tác kênh để tăng chi phí đầu thầu cho các hợp đồng chính phủ và để ngăn cản các nhà bán lại bán hộp mực và bột mực giả. Công ty này đã bị phạt 1,3 tỷ rupee vì hành vi này, và thêm 119,8 triệu rupee vì "tham gia vào cartelization" của bột mực, hộp mực và các vật tư in ấn khác. Hai mươi một nhà bán lại của HP cũng bị phạt 35,2 triệu rupee. CCI đã ra lệnh cho HP Ấn Độ và các đối tác kênh của họ phải chấm dứt và không được tiếp tục thực hiện các hành vi chống cạnh tranh và phải tổ chức các chương trình đào tạo tuân thủ cạnh tranh trong vòng 60 ngày. Điều tra đã tiết lộ rằng HP Ấn Độ đã đóng vai trò trung tâm trong việc sắp xếp thông đồng, nhằm mục đích cạnh tranh với các OEM khác và duy trì độc quyền trên thị trường.

7

GitHub cho người mới bắt đầu: Lộ trình để nắm vững các kiến thức cơ bản về GitHub

GitHub for Beginners: Your roadmap to mastering the GitHub essentials

GitHub Blog

[Đọc bài viết →](#)

GitHub đã tạo ra một hướng dẫn toàn diện cho người mới bắt đầu để□□ các yếu tố cơ bản của nền tảng. Hướng dẫn này bao gồm các chủ đề khác nhau, bao gồm trí tuệ nhân tạo và học máy trong hệ sinh thái

GitHub và ngành công nghiệp rộng lớn hơn. Nó khám phá các khả năng và lợi ích của việc tạo mã bằng AI và tác động của nó đối với trải nghiệm của nhà phát triển. Ngoài ra, nó cung cấp tài nguyên cho các nhà phát triển để phát triển kỹ năng và sự nghiệp, học cách xây dựng, vận hành và bảo trì phần mềm với GitHub, và khám phá công nghệ đằng sau nền tảng phát triển được hỗ trợ bởi AI hàng đầu thế giới. Hướng dẫn này cũng đi sâu vào các phương pháp hay nhất để xây dựng phần mềm với quy mô lớn với đội ngũ làm việc từ xa, bảo mật trong chu kỳ phát triển của nhà phát triển, và cách viết, xây dựng và triển khai phần mềm doanh nghiệp với quy mô lớn. Nó giải thích tại sao Gartner đã xếp GitHub vào vị trí Leader trong hai năm liên tiếp và giới thiệu việc tạo ra retrieval-augmented generation (RAG) để thu thập thêm thông tin. Đối với những người mới đến với GitHub, hướng dẫn này cung cấp một con đường chi tiết từ việc hiểu kiểm soát phiên bản, kho lưu trữ và yêu cầu kéo đến hợp tác trên các dự án thực tế và đóng góp cho mã nguồn mở. Nó nhằm mục đích cung cấp một mô hình hoàn chỉnh về cách phần mềm hiện đại được xây dựng trên GitHub, khiến nó trở thành một tài nguyên quý giá cho các nhà phát triển ở mọi cấp độ.

8

Tại sao thanh thiếu niên xứng đáng được tiếp cận AI an toàn

Why teens deserve access to safe AI

OpenAI Blog

[Đọc bài viết →](#)

OpenAI đang làm việc để làm cho nền tảng ChatGPT của mình an toàn hơn cho thanh thiếu niên bằng cách triển khai các biện pháp bảo vệ phù hợp với lứa tuổi. Công ty nhằm mục đích cung cấp một môi trường an toàn cho người dùng trẻ tương tác với công nghệ AI. Để đạt được điều này, OpenAI đang tích hợp các công cụ học tập và kiểm soát của phụ huynh vào nền tảng ChatGPT. Các tính năng này được thiết kế để giúp thanh thiếu niên học hỏi từ AI trong khi giảm thiểu các rủi ro tiềm năng. Ngoài ra, OpenAI đang hợp tác với các chuyên gia trong lĩnh vực này để nâng cao sự an toàn và hiệu quả của nền tảng cho người dùng trẻ. Bằng cách thực hiện các bước này, OpenAI nhằm mục đích cung cấp cho thanh thiếu niên quyền truy cập vào công nghệ AI trong khi đảm bảo sự an toàn và phúc lợi trực tuyến của họ. Các nỗ lực của công ty nhằm mục đích thúc đẩy một trải nghiệm lành mạnh và giáo dục cho người dùng trẻ, cho phép họ khám phá các khả năng của AI trong khi được bảo vệ khỏi các tác hại tiềm năng.

TIPS & TRICKS CHO DEV

Tối ưu hóa LangSmith

Vấn đề: Khó kiểm soát chi phí khi sử dụng AI tools.

Cách làm: Sử dụng LangSmith để phân tích và tối ưu hóa chi phí. Ví dụ, prompt "Optimize LangSmith costs" có thể giúp giảm thiểu chi phí.

Đánh giá: Hiệu quả trong việc giảm chi phí, nên dùng khi cần tối ưu hóa ngân sách.

Tích hợp Langfuse

Vấn đề: Khó tích hợp AI tools vào hệ thống hiện có.

Cách làm: Sử dụng Langfuse để tích hợp AI tools vào hệ thống. Ví dụ, lệnh `langfuse integrate` có thể giúp tích hợp nhanh chóng.

Đánh giá: Hiệu quả trong việc tích hợp, nên dùng khi cần kết nối hệ thống.

Giám sát Arize Phoenix

Vấn đề: Khó giám sát hiệu suất của AI models.

Cách làm: Sử dụng Arize Phoenix để giám sát và phân tích hiệu suất. Ví dụ, prompt "Analyze Arize Phoenix metrics" có thể giúp đánh giá hiệu suất.

Đánh giá: Hiệu quả trong việc giám sát, nên dùng khi cần đánh giá hiệu suất của AI models.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Để giảm thiểu chi phí và tăng hiệu suất khi sử dụng các mô hình ngôn ngữ lớn (LLM), các nhà phát triển cần biết cách tối ưu hóa chúng. Điều này giúp giảm tải tài nguyên và cải thiện hiệu suất ứng dụng. Việc tối ưu hóa LLM cũng giúp giảm thiểu chi phí vận hành và tăng tốc độ xử lý.

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 xử lý. Ngoài ra, việc triển khai mô hình trên các nền tảng edge AI hoặc sử dụng các công cụ như Ollama, LM Studio, llama.cpp cũng có thể giúp giảm chi phí và tăng hiệu suất.

4. Tip hoặc bước tiếp theo: Nên bắt đầu bằng việc đánh giá hiệu suất và chi phí hiện tại của mô hình LLM, sau đó áp dụng các kỹ thuật tối ưu hóa như fine-tuning, LoRA và triển khai trên edge AI để đạt được hiệu suất và chi phí tối ưu.

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

Nguồn: Google News · Groq AI