

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Thời gian là thầy giáo tốt nhất, nhưng thật đáng tiếc — nó giết chết tất cả học trò của nó."

— Hector Berlioz

Cuộc sống hữu hạn nhắc ta đừng lãng phí thời gian — hãy học hỏi và sống trọn vẹn từng ngày.

TIN TỨC NỔI BẬT

1

Tổng quan Agent Framework: Chọn Runtime phù hợp cho thực thi AI 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 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, điều này rất quan trọng cho việc thực thi trí tuệ nhân tạo (AI) trong môi trường doanh nghiệp. Các framework trình điều khiển đóng vai trò là xương sống cho các ứng dụng AI, cho phép chúng tương tác với môi trường xung quanh và đưa ra quyết định. Bài viết nhấn mạnh tầm quan trọng của việc chọn thời gian chạy phù hợp cho thực thi AI doanh nghiệp, vì nó có thể ảnh hưởng đáng kể đến hiệu suất, khả năng mở rộng và khả năng bảo trì của các ứng dụng AI. Nó thảo luận về các framework trình điều khiển khác nhau, bao gồm Rasa, Dialogflow và Microsoft Bot Framework, mỗi framework có điểm mạnh và điểm yếu riêng. Bài viết cũng đề cập đến các thách thức liên quan đến việc chọn framework trình điều khiển phù hợp, chẳng hạn như tích hợp với các hệ thống hiện có, quản lý dữ liệu và bảo mật. Nó nhấn mạnh nhu cầu đánh giá kỹ lưỡng các yếu tố này để đảm bảo rằng framework được chọn đáp ứng nhu cầu cụ thể của doanh nghiệp. Cuối cùng, bài viết nhằm cung cấp hướng dẫn cho các tổ chức muốn triển khai các giải pháp AI, nhấn mạnh tầm quan trọng của việc xem xét kỹ lưỡng khi chọn một framework trình điều khiển cho thực thi AI doanh nghiệp.

2

AWS open-source MCP Server cho Bedrock AgentCore, tinh gọn phát triển AI Agent

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

MarkTechPost

[Đọc bài viết →](#)

Amazon Web Services (AWS) đã mở nguồn một máy chủ MCP (Nền tảng Đa Đám mây) cho Bedrock AgentCore, một framework được thiết kế để đơn giản hóa việc phát triển đại lý AI. Động thái này nhằm mục đích đơn giản hóa quá trình tạo và triển khai các đại lý AI trên nhiều nền tảng đám mây khác nhau. Máy chủ MCP là một thành phần chính của Bedrock AgentCore, cho phép các nhà phát triển xây dựng và quản lý các đại lý AI có thể chạy trên nhiều môi trường đám mây. Bằng cách mở nguồn máy chủ MCP, AWS đang cung cấp cho cộng đồng nhà phát triển một giải pháp tiêu chuẩn hóa và mở nguồn cho việc phát triển đại lý AI. Việc mở nguồn máy chủ MCP này dự kiến sẽ đẩy nhanh việc phát triển các đại lý AI và thúc đẩy sự hợp tác giữa các nhà phát triển. Nó cũng phù hợp với mục tiêu của AWS trong việc thúc đẩy phát triển AI không phụ thuộc vào đám mây, cho phép các nhà phát triển xây dựng và triển khai các đại lý AI có thể tích hợp liền mạch với các nền tảng đám mây khác nhau. Máy chủ MCP mở nguồn hiện đã có sẵn cho các nhà phát triển truy cập và sử dụng, mở rộng thêm khả năng của Bedrock AgentCore và hệ sinh thái phát triển AI rộng lớn hơn.

3

Tăng tốc phát triển với Amazon Bedrock AgentCore MCP server | Trí tuệ nhân tạo

Accelerate development with the Amazon Bedrock AgentCore MCP server | Artificial Intelligence

Amazon Web Services (AWS)

[Đọc bài viết →](#)

Amazon Web Services (AWS) đã giới thiệu máy chủ Amazon Bedrock AgentCore MCP, được thiết kế để tăng tốc phát triển trong lĩnh vực trí tuệ nhân tạo (AI). Máy chủ AgentCore MCP là một thành phần chính của nền tảng Amazon Bedrock, cung cấp một bộ công cụ và dịch vụ để xây dựng, triển khai và quản lý các model AI. Máy chủ AgentCore MCP là một máy chủ hiệu suất cao cho phép các nhà phát triển chạy và quản lý các công việc AI một cách hiệu quả hơn. Nó cung cấp một môi trường có thể mở rộng và bảo mật để đào tạo và triển khai các model AI, cho phép các nhà phát triển tập trung vào việc xây dựng và cải thiện các model của họ mà không phải lo lắng về cơ sở hạ tầng cơ bản. Với máy chủ Bedrock AgentCore MCP, các nhà phát triển có thể

tăng tốc quá trình phát triển của mình bằng cách tận dụng các khả năng AI tiên tiến và cơ sở hạ tầng của AWS. Máy chủ này được thiết kế để làm việc liền mạch với các dịch vụ AWS khác, như SageMaker và Lake Formation, để cung cấp một nền tảng phát triển AI toàn diện. Bằng cách sử dụng máy chủ Bedrock AgentCore MCP, các nhà phát triển có thể xây dựng và triển khai các model AI nhanh hơn và hiệu quả hơn, thúc đẩy đổi mới và tăng trưởng kinh doanh.

4

15 Ghế văn phòng tốt nhất năm 2026—Đã thử nghiệm 70 mẫu để chọn

15 Best Office Chairs of 2026—We Tested 70 to Pick Them

Wired [Đọc bài viết →](#)

Tìm kiếm chiếc ghế văn phòng phù hợp là điều quan trọng để có một cuộc sống làm việc tại nhà thoải mái và tập trung. Một chiếc ghế tốt có thể tạo ra sự khác biệt đáng kể trong việc ngăn chặn các vấn đề về lưng và đau đớn không cần thiết. Đội ngũ WIRED Reviews đã dành hơn bảy năm để thử nghiệm và đánh giá ghế văn phòng nhằm giúp người tiêu dùng đưa ra quyết định thông minh. Sau khi thử nghiệm 70 chiếc ghế, đội ngũ đã biên soạn danh sách 15 chiếc ghế văn phòng tốt nhất năm 2026, đáp ứng các ngân sách và nhu cầu khác nhau. Lựa chọn hàng đầu cho hầu hết mọi người là Branch Ergonomic Chair Pro, cung cấp sự cân bằng giữa giá trị, tính năng, thoải mái và khả năng điều chỉnh với giá cả phải chăng là 499 đô la. Chiếc ghế này có tay vịn điều chỉnh, độ sâu ghế có thể tùy chỉnh và hệ thống hỗ trợ thắt lưng có thể được tùy chỉnh theo nhu cầu cá nhân. Các lựa chọn hàng đầu khác bao gồm Steelcase Gesture, có giá 1.510 đô la, và một chiếc ghế từ Staples có giá 169 đô la, vượt trội so với các thương hiệu Amazon rẻ hơn. Danh sách bao gồm nhiều lựa chọn cho các ngân sách, loại hình cơ thể và sở thích khác nhau. Đội ngũ WIRED đã mời nhiều người đánh giá với chiều cao và loại hình cơ thể khác nhau để thử nghiệm và chứng nhận sự thoải mái và độ bền của từng chiếc ghế. Danh sách đã được cập nhật để phản ánh những thay đổi trên thị trường, bao gồm cả việc thêm LiberNovo Omni Pro và loại bỏ các mặt hàng ngừng sản xuất.

5

Chúng tôi chơi bóng chày cũng đỉnh phết: Oklahoma vươn lên thống trị bóng chày đại học

We're Pretty Dang Good at This Baseball Thing: Oklahoma's Rise to College Baseball Supremacy

Chương trình bóng chày của Đại học Oklahoma đã trải qua một sự thăng tiến đáng chú ý lên ngôi vị tối cao của bóng chày đại học. Vào năm 2022, đội Sooners đã giành được danh hiệu thứ ba của Giải vô địch thế giới bóng chày nam (MCWS), củng cố vị trí của họ như một trong những chương trình hàng đầu dưới radar của cả nước. Sự thành công của đội có thể được quy cho sự kết hợp của các yếu tố, bao gồm việc tuyển dụng những tài năng hàng đầu, sự lãnh đạo của Huấn luyện viên trưởng Skip Johnson, và đầu tư vào các cơ sở hiện đại. Văn hóa đội và hóa học của đội Sooners cũng là chìa khóa cho sự thành công của họ, cho phép họ làm việc cùng nhau một cách liền mạch và đạt được những kết quả ấn tượng trên sân. Tình trạng dưới radar của Oklahoma giúp họ khác biệt so với các chương trình thành công khác, và khả năng của họ trong việc vượt qua thách thức và đạt được thành công là một minh chứng cho sự kiên cường và quyết tâm của họ. Sự thành công của đội Sooners có những ý nghĩa sâu rộng đối với tương lai của bóng chày đại học, và họ sẽ là một đội được theo dõi trong những năm tới.

6

Từ các dự án open source đình đám đến OpenAI

From open source hits to OpenAI

Changelog

[Đọc bài viết →](#)

Trong tập này, Max Stoiber từ OpenAI chia sẻ kinh nghiệm của mình trong ngành công nghệ. Ông suy ngẫm về các dự án mã nguồn mở, bao gồm cả những dự án phổ biến như react-boilerplate và styled-components, cũng như những dự án ít được biết đến. Stoiber cũng nói về sự tham gia của mình trong việc phát triển Spectrum, cuối cùng đã trở thành một phần của GitHub và giúp định hình GitHub Discussions. Ngoài ra, ông chia sẻ kinh nghiệm xây dựng Stellate, một bộ nhớ đệm GraphQL đã được Shopify và The Guild mua lại. Cuộc trò chuyện cũng đề cập đến việc phát triển các ứng dụng ChatGPT, mà Stoiber tin rằng cung cấp một bề mặt mới cho phần mềm. Tập này được tài trợ bởi một số công ty, bao gồm Coder, WorkOS, Notion, Fly.io và Changelog++.

7

Không dừng lại sớm: Case-folding source code với tốc độ bộ nhớ

Don't stop early: Case-folding source code at memory speed

GitHub Blog

[Đọc bài viết →](#)

Đội ngũ kỹ sư của GitHub đã phát triển một phương pháp để thực hiện các hoạt động gấp chữ (case-folding) với tốc độ bộ nhớ. Quá trình này rất quan trọng đối với công cụ tìm kiếm mã của họ, Blackbird, vốn lập chỉ mục hơn 180 triệu kho lưu trữ, tổng cộng hơn 480TB mã nguồn. Việc gấp chữ là cần thiết để xóa bỏ sự khác biệt về chữ hoa và chữ thường giữa các chuỗi, cho phép trả về kết quả tìm kiếm chính xác. Đội ngũ đã đạt được điều này bằng cách sử dụng vòng lặp không phân nhánh (branch-free loop) và số học không gian byte (byte-space arithmetic), cho phép họ gấp chữ mọi byte của tìm kiếm mã với tốc độ lớn hơn 45 GiB/s trên một lõi đơn. Tối ưu hóa này rất đáng kể, vì nó cải thiện hiệu suất tổng thể của công cụ tìm kiếm mã và nâng cao trải nghiệm người dùng. Bằng cách làm cho hoạt động này nhanh hơn, GitHub có thể xử lý hiệu quả quy mô lớn của cơ sở mã và cung cấp kết quả tìm kiếm chính xác hơn cho người dùng.

8

Thứ Sáu cùng Bob

Fridays With Bob

IEEE Spectrum

[Đọc bài viết →](#)

Robert N. Charette, một tác giả nổi tiếng của IEEE Spectrum và nhà sinh thái rủi ro, đã là một nhân vật quan trọng trong ấn phẩm này trong hơn 25 năm. Các công việc của Charette, bao gồm bài viết cơ bản "Tại sao Phần mềm Thất bại", đã được đọc và tham khảo rộng rãi trong các lớp kỹ thuật tại trường đại học. Ông đã viết hàng trăm bài viết cho IEEE Spectrum, bao gồm một blog phổ biến có tên "The Risk Factor", nơi ghi lại các thảm họa phần mềm và giành giải thưởng Jesse H. Neal vào năm 2016. Các đóng góp của Charette cho IEEE Spectrum đã rất rộng lớn, bao gồm các chủ đề như phát triển phần mềm, quản lý rủi ro và sự phổ biến ngày càng tăng của phần mềm trong cuộc sống của chúng ta. Ông cũng đã viết các bài viết sâu sắc về các chủ đề khác nhau, bao gồm hồ sơ sức khỏe điện tử, chuyển đổi xe điện và nghịch lý tự động hóa. Trong suốt sự nghiệp của mình, mục tiêu chính của Charette đã là làm cho phần mềm trở nên rõ ràng và dễ hiểu đối với công chúng. Ông tin rằng phần mềm thường vô hình, nhưng hậu quả của nó được mọi người cảm nhận. Charette hiện đã quyết định tập trung vào nhiếp ảnh thiên nhiên và viết một loạt truyện ba phần, bao gồm một phần về một kỹ sư trở thành thám tử.

TIPS & TRICKS CHO DEV

Chain-of-Thought Prompting

Vấn đề: Model không hiểu được logic và luồng suy nghĩ của dev.

Cách làm: Sử dụng chain-of-thought prompting, ví dụ: "Bước 1: Xác định vấn đề, Bước 2: Tìm kiếm giải pháp, Bước 3: Triển khai giải pháp".

Đánh giá: Hiệu quả khi cần giải quyết vấn đề phức tạp, nhưng yêu cầu dev phải có kỹ năng lập trình tốt.

Few-Shot Learning

Vấn đề: Model không thể học hỏi từ dữ liệu hạn chế.

Cách làm: Sử dụng few-shot learning, ví dụ: cung cấp 1-2 ví dụ cho model Claude về một task cụ thể.

Đánh giá: Hiệu quả khi dữ liệu hạn chế, nhưng yêu cầu dev phải có hiểu biết về task và model.

System Prompt Design

Vấn đề: Model không thể hiểu được yêu cầu của dev.

Cách làm: Sử dụng system prompt design, ví dụ: "Tạo một chương trình Python để giải quyết vấn đề này".

Đánh giá: Hiệu quả khi cần triển khai giải pháp cụ thể, nhưng yêu cầu dev phải có kỹ năng thiết kế hệ thống tốt.

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 Large Language Model (LLM). Việc này giúp giảm thiểu chi phí vận hành và cải thiện trải nghiệm người dùng. Các kỹ thuật như fine-tuning và LoRA có thể được áp dụng để đạt được mục tiêu này.

3. Ví dụ, sử dụng kỹ thuật LoRA (Low-Rank Adaptation) để điều chỉnh mô hình LLM cho một use case cụ thể, giúp giảm thiểu kích thước mô hình và tăng tốc độ xử lý.

4. Tip hoặc bước tiếp theo: Nghiên cứu và áp dụng các kỹ thuật như quantization, pruning và knowledge distillation để tối ưu hóa LLM, từ đó giảm chi phí và cải thiện hiệu năng cho ứng dụng AI của bạn.

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

Nguồn: Google News · Groq AI