

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Chữ tâm kia mới bằng ba chữ tài.”

— Nguyễn Du

Đức hạnh và tấm lòng tốt có giá trị hơn cả tài năng — người thành công thật sự phải có cả tâm lẫn tài.

TIN TỨC NỔI BẬT

1 Giới thiệu Open Agent Specification (Agent Spec): Một đặc tả thống nhất cho AI Agent

Introducing the Open Agent Specification (Agent Spec): A Unified Representation for AI Agents

Oracle Blogs

[Đọc bài viết →](#)

Oracle đã giới thiệu Open Agent Specification (Agent Spec), một biểu diễn thống nhất cho các tác nhân AI. Agent Spec nhằm cung cấp một khuôn khổ tiêu chuẩn hóa để mô tả và tương tác với các tác nhân AI, cho phép tích hợp và cộng tác liền mạch trên các hệ thống và nền tảng khác nhau. Thông số kỹ thuật này được thiết kế để mở và có thể mở rộng, cho phép các nhà phát triển dễ dàng thêm các tính năng và khả năng mới khi cần. Agent Spec được xây dựng trên các công nghệ hiện có như JSON và GraphQL, giúp nó dễ dàng tiếp cận và hiểu được đối với các nhà phát triển. Nó cung cấp một cách rõ ràng và concisely để định nghĩa hành vi, khả năng và tương tác của các tác nhân AI, giúp việc giao tiếp và khả năng tương tác giữa các hệ thống trở nên dễ dàng hơn. Bằng cách áp dụng Agent Spec, các nhà phát triển có thể tạo ra các ứng dụng AI phức tạp và tích hợp hơn, và các tổ chức có thể tận dụng lợi ích của AI một cách hiệu quả hơn. Agent Spec là một dự án mã nguồn mở, và Oracle mời cộng đồng nhà phát triển đóng góp vào sự phát triển và tiến hóa của nó.

2 Agent Factory: Kết nối các agent, app và dữ liệu bằng các chuẩn mở mới như MCP và A2A

Agent Factory: Connecting agents, apps, and data with new open standards like MCP and A2A

Microsoft đã giới thiệu Agent Factory, một công cụ được thiết kế để kết nối các agent, ứng dụng và dữ liệu bằng cách sử dụng các tiêu chuẩn mở mới. Agent Factory tận dụng Microsoft Cloud PC (MCP) và giao thức Application to Application (A2A) để tạo điều kiện cho các tương tác liền mạch giữa các hệ thống khác nhau. Tiêu chuẩn MCP cho phép kết nối an toàn và hiệu quả giữa các ứng dụng và dịch vụ dựa trên đám mây, trong khi giao thức A2A cho phép trao đổi dữ liệu hai chiều giữa các ứng dụng. Sự tích hợp này cho phép các nhà phát triển tạo ra các giải pháp mạnh mẽ và có khả năng mở rộng hơn. Bằng cách tận dụng các tiêu chuẩn mở này, các nhà phát triển có thể xây dựng các ứng dụng có thể giao tiếp với nhau và truy cập dữ liệu chung, nâng cao sự hợp tác và năng suất. Agent Factory được xây dựng trên Microsoft Azure, cung cấp một nền tảng có khả năng mở rộng và bảo mật cho các nhà phát triển để tạo và triển khai các ứng dụng của họ. Việc giới thiệu Agent Factory và các tiêu chuẩn mở liên quan nhằm mục đích đơn giản hóa quá trình tích hợp các hệ thống và ứng dụng khác nhau, cho phép các nhà phát triển tập trung vào việc xây dựng các giải pháp sáng tạo thay vì phải đối mặt với các vấn đề kết nối phức tạp.

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 Amazon Bedrock AgentCore MCP là một thành phần chính của nền tảng Amazon Bedrock, cung cấp một môi trường toàn diện để xây dựng, thử nghiệm và triển khai 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 khối lượng công việc AI một cách hiệu quả. Nó cung cấp một cơ sở hạ tầng có thể mở rộng và bảo mật để đào tạo, thử nghiệm 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. Máy chủ Amazon Bedrock AgentCore MCP được tích hợp với các dịch vụ AWS khác, chẳng hạn như SageMaker và Lake Formation, để cung cấp một trải nghiệm liền mạch cho các nhà phát triển. Sự tích hợp này cho phép

các nhà phát triển tận dụng điểm mạnh của từng dịch vụ để xây dựng và triển khai các model AI một cách hiệu quả hơn. Bằng cách sử dụng máy chủ Amazon Bedrock AgentCore MCP, các nhà phát triển có thể tăng tốc quá trình phát triển AI của họ, giảm chi phí và cải thiện hiệu suất tổng thể của các khối lượng công việc AI.

4

Liệu Bose có thể giúp Skullcandy rũ bỏ danh tiếng 'hàng giá rẻ'?

Can Bose Help Skullcandy Shake Its Bargain-Bin Reputation?

Wired [Đọc bài viết →](#)

Skullcandy, một công ty tai nghe nổi tiếng với các sản phẩm giá cả phải chăng nhưng thường bị chỉ trích vì chất lượng âm thanh không ấn tượng, đang nỗ lực cải thiện danh tiếng của mình. Công ty đã hợp tác với Bose để tạo ra các sản phẩm âm thanh chất lượng cao, bắt đầu với tai nghe Skullcandy Method 360 ANC và bây giờ là tai nghe Crusher 1080 ANC. Tai nghe Crusher mới vẫn giữ công nghệ tăng cường bass đặc trưng của thương hiệu, nhưng với chất lượng âm thanh được cải thiện nhờ chuyên môn của Bose. Tai nghe này có mảng microphone sáu chiếc để đảm bảo chất lượng cuộc gọi, âm thanh không gian và hủy tiếng ồn. Mặc dù là sản phẩm hàng đầu, Crusher 1080 ANC sẽ được bán lẻ với giá 280 đô la, một mức giá thấp hơn so với nhiều loại tai nghe cao cấp khác. Công ty tuyên bố tai nghe này cung cấp 60 giờ thời lượng pin và khả năng sạc nhanh. Với sản phẩm mới này, Skullcandy nhằm mục đích tái định vị mình như một thương hiệu lối sống độc đáo với các sản phẩm âm thanh chất lượng cao, chứ không chỉ là một lựa chọn tiết kiệm ngân sách.

5

Điều phối agent: Các tổ chức AI doanh nghiệp đang gặp vấn đề về triển khai, không phải về nền tảng — và hầu hết đang gọi chatbot là agent

Agentic orchestration: Enterprise AI organizations have a deployment problem, not a platform problem — and most are calling chatbots agents

VentureBeat [Đọc bài viết →](#)

Một cuộc khảo sát gần đây đối với 101 doanh nghiệp có 100 nhân viên hoặc hơn đã tiết lộ một khoảng cách giữa tham vọng và thực tế của việc điều phối đại lý AI doanh nghiệp. Phần lớn người được hỏi (40%) đang sử dụng Claude của Anthropic làm nền tảng điều phối đại lý chính, tiếp theo là Microsoft (18%) và OpenAI (13%). Sự lựa chọn nền

tăng được thúc đẩy bởi "trọng lực của mô hình" của mô hình cơ bản, với các doanh nghiệp ưu tiên sự phù hợp bản địa với mô hình cơ bản hiện đại. Tuy nhiên, mặc dù có tham vọng này, hầu hết các "đại lý" được triển khai vẫn chỉ là các trình bao chatbot một lời nhắc, với chỉ 10% người được hỏi đã vượt qua điểm giữa về mặt luồng công việc được điều phối nhiều bước thực sự. Các doanh nghiệp đang đầu tư vào công cụ và quyền cho luồng công việc, nhưng kiểm soát tài chính đối với việc tiêu thụ token của đại lý vẫn là một thách thức, với hơn một phần tư người được hỏi không có cách nào để dừng một đại lý bị lỗi trước khi hóa đơn đến. Cuộc khảo sát cũng cho thấy rằng các doanh nghiệp đang tập trung vào các nền tảng cung cấp mô hình trong khi đồng thời lên kế hoạch bao bọc chúng trong logic điều khiển tùy chỉnh, cho thấy mong muốn về một mặt phẳng điều khiển lai giữ kiểm soát bên ngoài nhà cung cấp. Sự khóa nhà cung cấp là mối quan ngại chính thúc đẩy cách tiếp cận này, với 35% người được hỏi cho rằng đó là rủi ro họ sợ nhất nếu kiểm soát nằm bên trong nhà cung cấp mô hình. Tổng thể, cuộc khảo sát nhấn mạnh nhu cầu của các doanh nghiệp trong việc bắc cầu khoảng cách giữa tham vọng và thực tế của việc điều phối, và phát triển các mặt phẳng điều khiển tinh vi hơn có thể quản lý sự phức tạp của luồng công việc nhiều bước và kiểm soát tài chính.

6

Những bài học về xây dựng agent từ quá trình phát triển Shippy

What building Shippy taught us about building agents

Hugging Face Blog [Đọc bài viết →](#)

Sự phát triển của Shippy, một đại lý AI hàng hải, đã cung cấp những bài học quý giá cho việc xây dựng các hệ thống AI đáng tin cậy và đáng. Shippy được thiết kế cho các quyết định có mức độ rủi ro cao, chẳng hạn như bảo vệ đại dương, nơi những câu trả lời sai có thể có hậu quả đáng kể. Đội ngũ Skylight đã ưu tiên xây dựng một hệ thống có thể được tin cậy để cung cấp thông tin chính xác, nằm trong giới hạn của nó và hoạt động nhất quán trên nhiều nhiệm vụ khác nhau. Họ đã đạt được điều này bằng cách chia đại lý thành ba thành phần: "linh hồn" (lời nhắc hệ thống), kỹ năng (xử lý yêu cầu cụ thể) và cấu hình (cài đặt thời gian chạy và phụ thuộc). Các kỹ năng được lưu trữ trong các tệp markdown đơn giản, giúp chúng dễ hiểu, có phiên bản và dễ sửa đổi. Các kỹ năng của Shippy có thể được kết hợp để trả lời các câu hỏi phức tạp, chẳng hạn như hiển thị hoạt động đánh bắt cá trong một khu vực cụ thể. Linh hồn của đại lý xác định ranh giới của

nó, đảm bảo nó không đưa ra quyết định pháp lý hoặc suy đoán vượt quá dữ liệu được hỗ trợ. Để làm cho hệ thống trở nên dự đoán hơn, Shippy sử dụng một CLI được thiết kế riêng để tương tác với API Skylight, thay vì tự xây dựng các cuộc gọi API. Cách tiếp cận này cho phép kiểm soát tốt hơn các công cụ mà đại lý sử dụng, khiến nó trở thành một hệ thống đáng tin cậy và đáng□ cho các lĩnh vực hoạt động có mức độ rủi ro cao.

7

xAI kiện một người đàn ông vì dùng Grok tạo 'deepfake' CSAM

xAI sues a man for using Grok to generate CSAM ‘deepfakes’

The Verge AI

[Đọc bài viết →](#)

Công ty xAI của Elon Musk đang kiện Terry Wayne Harwood, một người đàn ông đến từ Nam Carolina, vì cáo buộc sử dụng chatbot AI Grok của công ty để tạo ra tài liệu lạm dụng tình dục trẻ em (CSAM). xAI cho rằng Harwood đã vượt qua các biện pháp bảo vệ của chatbot để tạo và phân phối CSAM, vi phạm các chính sách của công ty. Harwood trước đó đã bị bắt vì sở hữu và phân phối CSAM và đang đối mặt với tám cáo buộc trọng tội. Vụ kiện cho biết một số hình ảnh liên quan đến các cáo buộc của Harwood đã được tạo hoặc thay đổi bằng cách sử dụng Grok. xAI đang tìm kiếm bồi thường về uy tín và pháp lý, cũng như bồi thường cho việc tự vệ khỏi các hành động pháp lý tiềm năng. Công ty cũng yêu cầu tòa án ra lệnh cấm Harwood tạo tài khoản xAI hoặc sử dụng Grok. Vụ kiện này diễn ra sau khi xAI phải đối mặt với sự chỉ trích vì thêm một tính năng vào Grok cho phép người dùng chỉnh sửa hình ảnh, dẫn đến việc tạo ra các AI deepfake, bao gồm cả những hình ảnh có người chưa thành niên.

8

Hình dạng, đối xứng và cấu trúc: Vai trò thay đổi của toán học trong nghiên cứu Machine Learning

Shape, Symmetries, and Structure: The Changing Role of Mathematics in Machine Learning Research

The Gradient

[Đọc bài viết →](#)

Trong thập kỷ qua, nghiên cứu học máy đã chuyển hướng sang các phương pháp đòi hỏi tính toán mạnh và dẫn dắt bởi kỹ thuật, dẫn đến những đột phá đáng kể thường vượt qua sự hiểu biết lý thuyết. Điều này đã dẫn đến sự suy đoán về vai trò giảm dần của toán học trong nghiên cứu học máy. Tuy nhiên, các tác giả cho rằng toán học vẫn còn □□ như bao giờ hết, vai trò của nó đang phát triển để giải quyết các

thách thức mới. Mặc dù toán học có thể không còn cung cấp cái nhìn sâu sắc ngay lập tức về những đột phá mới nhất, nó vẫn có thể được áp dụng theo nhiều cách, chẳng hạn như cung cấp lời giải thích hậu học về các hiện tượng kinh nghiệm và hướng dẫn các lựa chọn thiết kế cấp cao. Sự chuyển hướng sang quy mô cũng đã mở rộng phạm vi của các lĩnh vực toán học có thể áp dụng cho học máy, bao gồm cả các lĩnh vực toán học "thuần túy" như tô pô, đại số và hình học. Những lĩnh vực này đã phát triển khả năng xử lý các mức độ trừu tượng và phức tạp cao, hứa hẹn giải quyết các thách thức lớn trong học sâu. Bài viết sẽ khám phá các lĩnh vực nghiên cứu hiện tại chứng minh khả năng bền vững của toán học trong việc hướng dẫn khám phá và hiểu biết trong học máy, bao gồm cả các ứng dụng của toán học trong việc phát triển các mô hình AI, API, LLM mới và cải tiến các framework hiện có để hỗ trợ các nhà phát triển trong việc tạo ra các hệ thống học máy hiệu quả hơn, cũng như việc sử dụng token và các kỹ thuật khác để nâng cao hiệu suất của các mô hình.

TIPS & TRICKS CHO DEV

Local LLM với Ollama

Vấn đề: Phát triển ứng dụng cần chạy LLM mà không có internet.

Cách làm: Sử dụng Ollama, một công cụ hỗ trợ chạy Local LLM. Ví dụ, chạy lệnh `ollama --model lmstudio` để khởi động mô hình.

Đánh giá: Hiệu quả khi phát triển ứng dụng offline, nhưng cần cấu hình phù hợp.

LM Studio Integration

Vấn đề: Tích hợp LM Studio vào dự án hiện có.

Cách làm: Sử dụng API của LM Studio, ví dụ `lmstudio.load_model()` để tải mô hình.

Đánh giá: ☐ và hiệu quả khi tích hợp vào dự án.

Đánh giá Hiệu suất

Vấn đề: Đánh giá hiệu suất của Local LLM.

Cách làm: Sử dụng lệnh `ollama --benchmark` để đánh giá hiệu suất.

Đánh giá: Hiệu quả khi đánh giá và tối ưu hóa hiệu suất của mô hình.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Trong phát triển ứng dụng AI, việc tối ưu chi phí và hiệu năng của các mô hình ngôn ngữ lớn (LLM) là rất quan trọng. Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý, từ đó cải thiện trải nghiệm người dùng. Các dev cần biết cách

tối ưu hóa LLM để đảm bảo ứng dụng của họ hoạt động hiệu quả.

3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) để tối ưu hóa LLM cho use case cụ thể, giảm thiểu kích thước mô hình và tăng tốc độ xử lý.

4. Tip hoặc bước tiếp theo: Sử dụng các công cụ như Hugging Face Transformers để thực hiện fine-tuning và LoRA cho LLM, và đánh giá hiệu suất của mô hình sau khi tối ưu hóa.

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

Nguồn: Google News · Groq AI