BIS - MT · 14/05/2026

🤖 Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

✨ *"With great power comes great responsibility."*

↪ Quyền năng lớn đi kèm trách nhiệm lớn.

— Stan Lee (Spider-Man)

💡 *Khi nắm giữ công nghệ, kiến thức hay vị trí ảnh hưởng, trách nhiệm sử dụng chúng đúng đắn và có đạo đức là điều không thể thiếu.*

📰 TIN TỨC NỔI BẬT

1

Lớp lệnh AI đa agent cho kho hàng: Nâng tầm vận hành và thông tin chuỗi cung ứng

🇬🇧 *Multi-Agent Warehouse AI Command Layer Enables Operational Excellence and Supply Chain Intelligence*

📰 NVIDIA Developer 🔗 Đọc bài viết →

> NVIDIA đã giới thiệu Lớp lệnh AI Nhà kho đa tác nhân, một công nghệ tiên tiến được thiết kế để tối ưu hóa hoạt động nhà kho và quản lý chuỗi cung ứng. Giải pháp đổi mới này tận dụng trí tuệ nhân tạo (AI) để tự động hóa quy trình, nâng cao hiệu quả và cung cấp thông tin theo thời gian thực về hiệu suất chuỗi cung ứng. Lớp lệnh AI Nhà kho đa tác nhân cho phép giao tiếp và phối hợp liền mạch giữa các bên liên quan khác nhau, bao gồm robot, máy bay không người lái và công nhân người. Bằng cách tích hợp việc ra quyết định dựa trên AI, hệ thống có thể tối ưu hóa các nhiệm vụ như quản lý hàng tồn kho, thực hiện đơn hàng và lập kế hoạch hậu cần. Lợi ích chính của công nghệ này bao gồm hiệu quả hoạt động được cải thiện, giảm chi phí và tăng khả năng hiển thị chuỗi cung ứng. Lớp lệnh AI cung cấp phân tích dữ liệu theo thời gian thực, cho phép doanh nghiệp đưa ra quyết định thông minh và phản ứng nhanh chóng với những thay đổi trong nhu cầu hoặc cung cấp. Bằng cách triển khai công nghệ này, các tổ chức có thể đạt được sự xuất sắc trong hoạt động, cải thiện sự hài lòng của khách hàng và duy trì tính cạnh tranh trên thị trường nhanh chóng ngày nay. Lớp lệnh AI Nhà kho đa tác nhân là một bước tiến quan trọng trong việc phát triển các giải pháp quản lý chuỗi cung ứng dựa trên AI.

2 Agent Factory: Kết nối các agent, app và data với các open standard mới như MCP và A2A

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

📰 Microsoft Azure 🔗 Đọc bài viết →

Microsoft đã giới thiệu Agent Factory, một nền tảng nhằm kết nối các tác nhân, ứng dụng và dữ liệu thông qua các tiêu chuẩn mở mới. Nền tảng này tận dụng Nền tảng Điện toán Đám mây (MCP) và Tiêu chuẩn Ứng dụng-sang-Ứng dụng (A2A) của Microsoft. Agent Factory cho phép tương tác liền mạch giữa các thành phần khác nhau, chẳng hạn như tác nhân, ứng dụng và nguồn dữ liệu, bằng cách cung cấp một giao diện tiêu chuẩn hóa. Điều này giúp tạo ra các hệ thống hiệu quả và có khả năng mở rộng hơn. Các tiêu chuẩn mở của nền tảng cho phép linh hoạt và khả năng tương tác cao hơn, giúp dễ dàng tích hợp các công cụ và dịch vụ khác nhau. Bằng cách sử dụng Agent Factory, các nhà phát triển có thể xây dựng các hệ thống phức tạp và động hơn, có thể thích nghi với các yêu cầu thay đổi. Các tiêu chuẩn mở của nền tảng cũng thúc đẩy sự hợp tác và đổi mới trong cộng đồng nhà phát triển. Việc Microsoft giới thiệu Agent Factory là một bước quan trọng hướng tới tạo ra một hệ sinh thái công nghệ kết nối và hiệu quả hơn. Nền tảng này dự kiến sẽ có tác động tích cực đến sự phát triển của các ứng dụng và hệ thống trong tương lai.

3 Giới thiệu Open Agent Specification (Agent Spec): Một cách biểu diễn thống nhất cho các 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 chung cho việc phát triển và tích hợp các tác nhân AI trên các nền tảng và ứng dụng khác nhau. Bằng cách tiêu chuẩn hóa biểu diễn của các tác nhân AI, Agent Spec tìm cách tạo điều kiện cho tính tương tác và khả năng di chuyển của các model AI, cho phép tích hợp liền mạch với các hệ thống và công cụ khác nhau. Agent Spec được thiết kế để mở và có thể mở rộng, cho phép các nhà phát triển tạo ra các tác nhân tùy chỉnh có thể tương tác với các hệ thống và ứng dụng khác nhau. Tiêu chuẩn này dự kiến sẽ thúc đẩy sự phát triển của các giải pháp AI tinh vi và tích hợp hơn, thúc đẩy đổi mới trong các lĩnh

vực như AI đối thoại, robot và hệ thống tự động. Việc giới thiệu Agent Spec đại diện cho một bước tiến quan trọng hướng tới việc tạo ra một hệ sinh thái AI thống nhất và kết nối hơn, nơi các tác nhân AI khác nhau có thể làm việc cùng nhau một cách liền mạch và hiệu quả. Khi cảnh quan AI tiếp tục phát triển, Agent Spec có vị thế để đóng vai trò quan trọng trong việc định hình tương lai của phát triển và triển khai AI.

4

AI chatbot đang tiết lộ số điện thoại thật của người dùng

🇬🇧 *AI chatbots are giving out people's real phone numbers*

📰 MIT Tech Review 🔗 Đọc bài viết →

Trợ lý trò chuyện AI của Google, Gemini, đã làm lộ số điện thoại thực của mọi người, khiến người dùng dễ bị nhận các cuộc gọi không mong muốn và có thể bị quấy rối. Một số cá nhân đã báo cáo rằng thông tin liên hệ cá nhân của họ đã được trợ lý trò chuyện này tiết lộ, thường là do hướng dẫn hoặc phản hồi dịch vụ khách hàng không chính xác. Các chuyên gia cho rằng vấn đề này là do thông tin nhận dạng cá nhân (PII) được sử dụng trong dữ liệu đào tạo, mặc dù cơ chế chính xác vẫn chưa rõ ràng. Những sự việc này không bị cô lập, với sự tăng đáng kể các truy vấn của khách hàng về AI tạo sinh từ công ty DeleteMe. Cụ thể, 55% các mối quan tâm này đề cập đến ChatGPT, 20% đề cập đến Gemini và 15% đề cập đến Claude. Người dùng đã báo cáo nhận được thông tin cá nhân chính xác, bao gồm địa chỉ nhà, số điện thoại và chi tiết nhà tuyển dụng, khi đặt câu hỏi vô hại cho các trợ lý trò chuyện. Các chuyên gia cảnh báo rằng đây có thể là một vấn đề phổ biến, với nhiều trường hợp không được báo cáo. Sự thiếu kiểm soát đối với việc lộ dữ liệu cá nhân bởi các trợ lý trò chuyện AI làm dấy lên lo ngại về quyền riêng tư trực tuyến và khả năng bị quấy rối hoặc tương tác tiêu cực khác.

5

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

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

📰 The Gradient 🔗 Đọc bài viết →

Trong những năm gần đây, nghiên cứu học máy đã chứng kiến một sự thay đổi đáng kể, với tiến bộ thực nghiệm vượt qua sự hiểu biết lý thuyết. Điều này đã dẫn đến sự suy đoán về vai trò giảm dần của toán

học trong lĩnh vực này. Tuy nhiên, các nhà nghiên cứu cho rằng toán học vẫn còn□□ như bao giờ hết, vai trò của nó đơn giản là đang phát triển. Trong khi toán học có thể không còn cung cấp cái nhìn sâu sắc ngay lập tức về những đột phá, nó đang được sử dụng theo những cách mới, chẳng hạn như cung cấp lời giải thích hậu hoc về các hiện tượng thực nghiệm và hướng dẫn các lựa chọn thiết kế cấp cao. Sự quan trọng ngày càng tăng của quy mô trong học máy cũng đã mở rộng phạm vi của các lĩnh vực toán học có thể áp dụng cho lĩnh vực này, bao gồm cả các lĩnh vực toán học "tinh khiết" như tô pô, đại số và hình học. Những lĩnh vực này đang được sử dụng để giải quyết các thách thức phức tạp trong học sâu, chẳng hạn như hiểu hành vi của các mạng nơ-ron lớn. Bài viết khám phá các lĩnh vực nghiên cứu hiện tại mà thể hiện khả năng bền vững của toán học trong việc hướng dẫn khám phá và hiểu biết trong học máy, đặc biệt là trong việc phát triển các model AI, API, LLM mới và cải tiến các framework hiện có để hỗ trợ các developer trong việc tạo ra các ứng dụng thông minh hơn bằng cách sử dụng token và các kỹ thuật khác.

6

Managed Deep Agents: Cách nhanh nhất để triển khai một production deep agent

🇬🇧 *Managed Deep Agents: the fastest way to ship a production deep agent*

📰 LangChain Blog 🔗 Đọc bài viết →

Managed Deep Agents là một thời gian chạy được lưu trữ đầu tiên API mới được giới thiệu trong beta riêng, được thiết kế để đơn giản hóa quá trình vận chuyển các tác nhân sâu sản xuất. Nó cung cấp một ngôi nhà bền vững cho các tác nhân sâu trong LangSmith, cho phép các nhà phát triển tạo, cập nhật, quản lý và chạy các tác nhân một cách lập trình từ ứng dụng hoặc luồng công việc nền tảng nội bộ của riêng họ. Giải pháp này đóng gói lớp hoạt động xung quanh khung Deep Agents mã nguồn mở, cho phép các nhà phát triển tập trung vào hành vi của tác nhân thay vì xây dựng lại thời gian chạy. Managed Deep Agents hỗ trợ các luồng bền vững, chạy luồng, checkpointing và các công việc có sự tham gia của con người. Nó lưu trữ và phiên bản các tệp định nghĩa tác nhân trong LangSmith, cho phép tác nhân phát triển theo thời gian. Tính năng Context Hub cung cấp một vị trí được quản lý để các tác nhân giữ và cập nhật ngữ cảnh trên các lần chạy, trong khi LangSmith Engine xem xét các dấu vết của tác nhân để xác định các lỗi và khu vực cần cải thiện. Giải pháp này cũng hỗ trợ thực thi được hỗ trợ bởi sandbox cho các công việc yêu cầu mã, lệnh shell và I/O tệp. Các lần chạy được tự động theo dõi trong LangSmith, cung

cấp cho các nhà phát triển cùng một luồng quan sát mà họ sử dụng cho các tác nhân và ứng dụng LLM. Đường dẫn ra mắt là API-first, với beta riêng tập trung vào một tập hợp nhỏ các nguyên thủy được quản lý để chạy các tác nhân sâu trong LangSmith.

7

Xây dựng một sandbox an toàn, hiệu quả để kích hoạt Codex trên Windows

🇬🇧 *Building a safe, effective sandbox to enable Codex on Windows*

📰 OpenAI Blog 🔗 Đọc bài viết →

OpenAI đã phát triển một môi trường sandbox bảo mật để cho phép sử dụng Codex trên Windows. Môi trường sandbox này cung cấp một không gian an toàn và được kiểm soát để Codex, một tác nhân mã hóa, hoạt động trong đó. Môi trường sandbox đảm bảo rằng Codex chỉ có quyền truy cập hạn chế vào các tệp và khả năng mạng bị hạn chế, ngăn chặn các rủi ro bảo mật tiềm ẩn và các hành động không được ủy quyền. Bằng cách cô lập Codex trong môi trường sandbox, OpenAI đã tạo ra một môi trường bảo mật cho phép tác nhân mã hóa hoạt động hiệu quả trong khi giảm thiểu rủi ro về vi phạm dữ liệu hoặc thỏa hiệp hệ thống. Cách tiếp cận này cho phép các nhà phát triển sử dụng Codex trên Windows một cách tự tin, biết rằng hệ thống và dữ liệu của họ được bảo vệ. Môi trường sandbox là một thành phần quan trọng trong việc triển khai Codex trên Windows, vì nó giải quyết các mối quan ngại về bảo mật và bảo vệ dữ liệu. Bằng cách cung cấp một không gian được kiểm soát và cô lập cho Codex hoạt động, OpenAI đã làm cho nó có thể cho các nhà phát triển tận dụng các khả năng của Codex trong khi duy trì mức độ bảo mật hệ thống cao.

8

AI IQ đã ra mắt: Trang web mới chấm điểm các frontier AI model theo thang IQ của con người. Kết quả đã gây chia rẽ trong giới tech.

🇬🇧 *AI IQ is here: a new site scores frontier AI models on the human IQ scale. The results are already dividing tech.*

📰 VentureBeat 🔗 Đọc bài viết →

Một trang web mới, AI IQ, đã được ra mắt để đánh giá các mô hình AI tiên phong trên thang điểm chỉ số thông minh của con người, gây ra cả lời khen ngợi và chỉ trích từ cộng đồng công nghệ. Trang web, được tạo bởi kỹ sư và doanh nhân Ryan Shea, phân bổ ước tính chỉ số thông minh cho hơn 50 mô hình ngôn ngữ mạnh mẽ nhất trên thế giới và vẽ

> chúng trên một đường cong hình chuông tiêu chuẩn. Kết quả cho thấy mô hình GPT-5.5 của OpenAI dẫn đầu đường cong hình chuông với ước tính chỉ số thông minh là 136, theo là các mô hình hàng đầu khác. Tuy nhiên, các nhà chỉ trích cho rằng việc giảm khả năng phức tạp của một mô hình ngôn ngữ xuống một con số duy nhất tạo ra một ảo tưởng sai lầm về độ chính xác và che giấu vấn đề "jaggedness", nơi các mô hình thể hiện khả năng không đồng đều. AI IQ cũng bao gồm một điểm số trí tuệ cảm xúc (EQ), tạo ra một xếp hạng khác với chỉ số IQ alone. Phương pháp luận của trang web dựa trên một công thức nhóm 12 điểm chuẩn vào bốn chiều hướng lý luận: trừu tượng, toán học, lập trình và học thuật. Chỉ số thông minh tổng hợp là trung bình trực tiếp của bốn chiều hướng điểm số này. Các biểu đồ của trang web đã được các nhà công nghệ doanh nghiệp khen ngợi vì làm cho thị trường phức tạp trở nên dễ hiểu, nhưng các nhà chỉ trích cho rằng phương pháp luận là một phần không rõ ràng và rằng ẩn dụ IQ có thể gây hiểu lầm. Mặc dù có những chỉ trích này, AI IQ cung cấp một khuôn khổ duy nhất để so sánh các mô hình trên các nhà cung cấp, chiều hướng, và điểm giá, được cập nhật thường xuyên, với đủ sắc thái để chỉ ra rằng câu trả lời đúng cho "mô hình nào là tốt nhất?" thường là "tùy thuộc vào nhiệm vụ". Trang web này cho phép các nhà phát triển đánh giá và so sánh các mô hình LLM khác nhau thông qua một API, và cũng cung cấp một framework để xây dựng và đào tạo các mô hình AI mới.

⚡ TIPS & TRICKS CHO DEV

⚡ Chain-of-Thought Prompting

Vấn đề: Model AI không hiểu rõ context và mối quan hệ giữa các ý tưởng.

Cách làm: Sử dụng kỹ thuật chain-of-thought bằng cách đưa ra một chuỗi các câu hỏi và hướng dẫn cụ thể, như "Bước 1: Xác định vấn đề, Bước 2: Phân tích dữ liệu, Bước 3: Kết luận".

Đánh giá: Hiệu quả trong việc tăng cường khả năng hiểu và phân tích của model AI, đặc biệt khi giải quyết vấn đề phức tạp.

⚡ Few-Shot Learning

Vấn đề: Model AI cần nhiều dữ liệu để huấn luyện và cải thiện hiệu suất.

Cách làm: Sử dụng kỹ thuật few-shot learning bằng cách cung cấp một số ví dụ hạn chế và hướng dẫn model AI học hỏi từ đó, như "Hãy viết một câu chuyện dựa trên 3 từ khóa: tình yêu, Paris, mùa xuân".

Đánh giá: Hiệu quả trong việc giảm thiểu nhu cầu dữ liệu huấn luyện, nhưng có thể không mang lại kết quả chính xác nếu không đủ thông tin.

⚡ System Prompt Design

Vấn đề: Model AI không hiểu rõ vai trò và chức năng của nó trong hệ thống.
Cách làm: Thiết kế một hệ thống prompt cụ thể, bao gồm cả vai trò và chức năng của model AI, như "Tôi là một trợ lý ảo, hãy giúp tôi trả lời câu hỏi của người dùng".
Đánh giá: Hiệu quả trong việc tăng cường khả năng hiểu và tương tác của model AI, đặc biệt khi được tích hợp vào các hệ thống phức tạp.

📚 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 mô hình ngôn ngữ lớn (LLM), các nhà phát triển cần biết cách tinh chỉnh và tối ưu hóa mô hình. Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý.
3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) để tinh chỉnh mô hình cho các use case cụ thể, giúp giảm thiểu số lượng tham số cần thiết và tăng tốc độ xử lý.
4. 💡 Tip hoặc bước tiếp theo: Nên bắt đầu bằng cách phân tích yêu cầu cụ thể của dự án và chọn mô hình LLM phù hợp, sau đó áp dụng kỹ thuật fine-tuning và LoRA để tối ưu hóa hiệu năng và giảm thiểu chi phí.