BIS - MT · 24/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Lời nói chẳng mất tiền mua, lựa lời mà nói cho vừa lòng nhau."

— Ca dao Việt Nam

Ngôn từ có sức mạnh lớn — một lời nói khéo léo, chân thành có thể xây dựng mối quan hệ tốt đẹp mà không tốn gì.

TIN TỨC NỔI BẬT

1 Giới thiệu Open Agent Specification (Agent Spec): Một chuẩn biểu diễn 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 chung cho các nhà phát triển để tạo, tích hợp và tương tác với các tác nhân AI khác nhau trên các nền tảng và lĩnh vực khác nhau. Đặc tả 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 vào các tác nhân AI. Agent Spec được xây dựng trên các công nghệ tiêu chuẩn của ngành như JSON và HTTP, giúp nó dễ dàng truy cập và tích hợp với các hệ thống hiện có. Nó cung cấp một cách tiêu chuẩn hóa để mô tả các tác nhân AI, bao gồm khả năng, ý định và tương tác của chúng. Điều này cho phép giao tiếp và cộng tác liền mạch giữa các tác nhân AI khác nhau, cũng như giữa các tác nhân AI và người dùng con người. Bằng cách áp dụng Agent Spec, các nhà phát triển có thể tạo ra các hệ thống AI phức tạp và tích hợp hơn có thể tương tác với nhau và với con người một cách có ý nghĩa 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 agent, ứng dụng và dữ liệu với các chuẩn mở 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 cho phép kết nối liền mạch giữa các tác nhân, ứng dụng và dữ liệu. Đổi mới này được xây dựng trên các tiêu chuẩn mở mới, bao gồm MCP (Nền tảng Kết nối Microsoft) và A2A (Ứng dụng đến Ứng dụng). Agent Factory nhằm mục đích đơn giản hóa tích hợp và giao tiếp giữa các hệ thống và dịch vụ khác nhau. Bằng cách tận dụng MCP và A2A, Agent Factory tạo điều kiện cho việc tạo ra các tác nhân thông minh có thể tương tác với các ứng dụng và nguồn dữ liệu khác nhau. Điều này cho phép có các quy trình làm việc hiệu quả và tự động hơn, cũng như trao đổi và phân tích dữ liệu được cải thiện. Các tiêu chuẩn mở của nền tảng đảm bảo rằng nó có thể thích ứng với các môi trường đa dạng và có thể dễ dàng tích hợp với các hệ thống hiện có. Agent Factory có tiềm năng cách mạng hóa cách các doanh nghiệp và tổ chức tương tác với các ứng dụng và dữ liệu của họ. Bằng cách cung cấp một khuôn khổ tiêu chuẩn hóa cho giao tiếp và tích hợp, nó có thể giúp tối ưu hóa quy trình, cải thiện năng suất và mở khóa những thông tin mới từ dữ liệu. Các chi tiết thêm về Agent Factory và khả năng của nó dự kiến sẽ được Microsoft phát hành trong những ngày tới.

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 dịch vụ quản lý để xây dựng, triển khai và quản lý các mô hình AI quy mô lớn. Máy chủ AgentCore MCP là một máy chủ hiệu suất cao cho phép các nhà phát triển đào tạo và triển khai các mô hình AI một cách hiệu quả hơn. Nó cung cấp một môi trường có thể mở rộng và bảo mật cho việc phát triển mô hình, cho phép các nhà phát triển tập trung vào việc xây dựng và tinh chỉnh các mô hình AI của họ mà không cần phải lo lắng về cơ sở hạ tầng cơ bản. Với máy chủ AgentCore MCP, các nhà phát triển có thể tận dụng các khả năng AI và học máy rộng lớn của AWS, bao gồm hỗ trợ cho các framework phổ biến như TensorFlow và PyTorch. Máy chủ

này cũng tích hợp với các dịch vụ AWS khác, như SageMaker và Lake Formation, để cung cấp một nền tảng phát triển AI toàn diện. Bằng cách tận dụng máy chủ Amazon Bedrock AgentCore MCP, các nhà phát triển có thể tăng tốc quá trình phát triển và triển khai AI của họ, cho phép họ đưa ra các giải pháp AI sáng tạo ra thị trường nhanh hơn.

4

LangChain Newsletter tháng 7/2026 — NemoClaw Blueprint, OpenWiki Brains và nhiều hơn nữa

July 2026: LangChain Newsletter — NemoClaw Blueprint, OpenWiki Brains, and More

LangChain Blog Đọc bài viết →

Trong Bản tin LangChain tháng 7 năm 2026, Jensen Huang và Harrison thảo luận về tầm quan trọng của các hệ thống tác nhân mở đối với các công ty. NVIDIA đã phát hành một bản thiết kế, NemoClaw, là một phần của Deep Agents của LangChain. Bản thiết kế này nhằm mục đích cho phép các đội xây dựng các hệ thống tác nhân mà họ có thể kiểm soát, điều chỉnh và sở hữu. Cộng đồng LangChain cũng đang mở rộng với hội nghị Interrupt sắp tới, sẽ được tổ chức tại Thành phố New York và London vào mùa thu này. Các lãnh đạo ngành, bao gồm Harrison Chase, sẽ phát biểu tại sự kiện, và người tham dự có thể tham gia các buổi hội thảo thực hành về xây dựng và gỡ lỗi tác nhân. Ngoài ra, LangSmith đã giới thiệu một số tính năng mới, bao gồm một phiên bản thử nghiệm miễn phí cho Sandboxes, cung cấp một môi trường an toàn để các tác nhân chạy mã và thử nghiệm thay đổi. LangSmith Fleet cho phép tích hợp tác nhân vào Slack, và việc theo dõi tác nhân giọng nói đã được cải thiện để thu thập các chỉ số âm thanh và độ trễ. OpenWiki Brains là một lớp bộ nhớ mới chuyển đổi các nguồn khác nhau thành một wiki cục bộ cho các tác nhân sử dụng. Bản tin cũng nhấn mạnh thành công của bộ khung tác nhân mã nguồn mở của LangChain, Deep Agents, trong việc xây dựng các tác nhân chạy dài cho các quy trình công việc phức tạp. Một số sự kiện sắp tới, bao gồm các cuộc gặp gỡ và LangSmith Roadshow, được công bố, và hai công ty, Schneider Electric và Pendo, chia sẻ kinh nghiệm của họ khi sử dụng nền tảng của LangChain, bao gồm việc sử dụng AI, API, LLM, model, token, developer, framework, v.v.

5 Microsoft ra mắt các AI model "cây nhà lá vườn" mới, tuyên bố giảm chi phí tới 89% so với OpenAI

Microsoft launches new in-house AI models it says cut costs up to 89% versus OpenAI

VentureBeat Đọc bài viết →

Microsoft đã ra mắt hai mô hình AI nội bộ mới, MAI-Image-2.5-Pro và MAI-Voice-2-Flash, vào phiên bản xem trước công khai. Những mô hình này được thiết kế để cắt giảm chi phí cho các sản phẩm và dịch vụ của Microsoft, với MAI-Image-2.5-Pro nhắm vào việc tạo hình ảnh cao cấp và MAI-Voice-2-Flash tập trung vào các ứng dụng giọng nói với khối lượng lớn. Công ty cho biết rằng các mô hình nội bộ của họ có thể giảm chi phí GPU lên đến 89% so với các mô hình frontier của OpenAI. Microsoft cũng đã xuất bản dữ liệu sản xuất cho thấy việc triển khai các mô hình nội bộ của họ trong các sản phẩm khác nhau, bao gồm Bing, PowerPoint, OneDrive, Dynamics 365 và Excel. Công ty báo cáo về việc tiết kiệm chi phí đáng kể và cải thiện hiệu suất, chẳng hạn như tăng 26% tỷ lệ lưu và giảm 25% độ trễ P95 trong OneDrive. Chiến lược của Microsoft bao gồm xây dựng các họ mô hình được tùy chỉnh cho các trường hợp sử dụng cụ thể, thay vì dựa vào một mô hình hàng đầu duy nhất. CEO của công ty, Satya Nadella, đã định hình cách tiếp cận này là "sự khuếch tán frontier", nơi các khả năng frontier bão hòa được cung cấp ở quy mô lớn và chi phí thấp hơn thông qua các mô hình được tối ưu hóa cho các sản phẩm sử dụng cao. Microsoft cũng đang đóng gói cuốn sách hướng dẫn AI nội bộ của mình thành một sản phẩm Azure, gọi là Frontier Tuning, cho phép các doanh nghiệp đào tạo các mô hình chuyên dụng chống lại các đánh giá và môi trường học tăng cường độc quyền của riêng họ. Điều này được xem là một cách để Microsoft kiếm tiền từ chuyên môn AI của mình và thu hút khách hàng doanh nghiệp đến với các dịch vụ đám mây của mình.

6 Ghép thận siêu lạnh thành công trên lợn, đánh dấu "thành tựu mang tính bước ngoặt"

Supercooled kidneys have been transplanted into pigs in a “landmark achievement”

MIT Tech Review Đọc bài viết →

Các nhà khoa học tại Đại học Texas A&M đã đạt được một bước tiến quan trọng trong bảo tồn cơ quan, thành công trong việc cấy ghép thận siêu lạnh vào lợn. Đội ngũ do Matthew Powell Palm dẫn đầu đã phát triển một thiết bị cho phép làm mát các cơ quan xuống -4°C (25°F) mà không hình thành băng, điều mà trước đây đã gây ra thiệt

hại. Phương pháp này, được gọi là siêu lạnh, cho phép các cơ quan được lưu trữ trong vài ngày thay vì vài giờ. Thiết bị sử dụng một thùng chứa điều khiển áp suất để duy trì nhiệt độ không đổi, ngăn chặn sự hình thành băng. Trong nghiên cứu, thận lợn được làm mát xuống -4°C và lưu trữ lên đến 72 giờ trước khi được cấy ghép thành công trở lại vào lợn cho ban đầu. Kết quả cho thấy thận siêu lạnh hoạt động tốt hơn những thận được lưu trữ trên băng. Thành tựu này có tiềm năng giúp giảm thiểu khủng hoảng thiếu hụt cơ quan, vì nó có thể cho phép lưu trữ lâu dài hơn các cơ quan người hiến tặng. Hiện tại, hơn 104.000 người ở Mỹ đang chờ đợi một ca cấy ghép thận, và 17 người tử vong mỗi ngày trong khi chờ đợi một ca cấy ghép.

7

Doanh nghiệp nhỏ ở California đón nhận ChatGPT: Cách tiếp cận thực tế để tích hợp AI

California Small Businesses Embrace ChatGPT: A Hands-On Approach to AI Integration

Dev.to AI Đọc bài viết →

Các doanh nghiệp nhỏ tại California đang tận dụng sức mạnh của trí tuệ nhân tạo, đặc biệt là ChatGPT, để tăng cường hiệu quả và đổi mới. OpenAI gần đây đã tổ chức một sự kiện tại Trung tâm Cộng đồng Tula, dành riêng cho các doanh nhân California, để cung cấp trải nghiệm thực tế với ChatGPT. Sự kiện này nhằm mục đích trao quyền cho các chủ doanh nghiệp tích hợp AI vào hoạt động hàng ngày của họ và mở khóa các hiệu quả mới. Traci Lee từ Bộ phận Chính phủ Tiểu bang và Địa phương của OpenAI đã chào đón các khách mời, nhấn mạnh các kết quả thực tế và có thể hành động được từ phiên họp. Thiết kế của sự kiện ưu tiên sự tương tác trực tiếp, cho phép người tham gia thử nghiệm với ChatGPT dưới sự hướng dẫn của chuyên gia và khám phá cách công cụ này có thể giải quyết các nhu cầu kinh doanh cụ thể của họ. Các khách mời ca ngợi cách tiếp cận thực tế, cho rằng nó hiệu quả trong việc□ dưỡng sự hiểu biết sâu sắc hơn và tính ứng dụng ngay lập tức của các khái niệm đã học. Các chuyên gia, bao gồm Rutger Hage và Hilda Soto, đã chia sẻ quan điểm của họ về tác động chuyển đổi của AI, nhấn mạnh lợi ích của nó cho các doanh nghiệp nhỏ và tầm quan trọng của việc đưa giáo dục AI đến cộng đồng rộng lớn hơn. Sự kiện này đã giúp người tham gia có cái nhìn sâu sắc về tiềm năng của ChatGPT trong việc tăng cường năng suất, tạo ra ý tưởng đổi mới và giải quyết các thách thức kinh doanh phức tạp, đảm bảo rằng các tham gia rời đi với những hiểu biết có thể hành động và một bản đồ rõ ràng hơn để triển khai các giải pháp AI trong chính doanh nghiệp của họ.

8

Code Finder: Tìm kiếm code nhanh chóng, hiệu quả cho coding agent

Code Finder: fast, efficient code search for coding agents

Sourcegraph Blog Đọc bài viết →

Code Finder là một công cụ được thiết kế để tối ưu hóa quá trình tìm kiếm mã cho các tác nhân mã hóa. Nó hoạt động bằng cách chạy vòng lặp tìm kiếm của riêng mình, cho phép nó tìm kiếm mã cần thiết một cách hiệu quả. Cách tiếp cận này cho phép Code Finder cung cấp cho tác nhân mã hóa các tệp và phạm vi dòng chính xác cần thiết, giảm đáng kể thời gian và chi phí liên quan đến quá trình tìm kiếm. Bằng cách tự thực hiện nhiệm vụ tìm kiếm, Code Finder loại bỏ nhu cầu của tác nhân mã hóa phải thực hiện tìm kiếm, dẫn đến một giải pháp nhanh hơn và tiết kiệm chi phí hơn. Chức năng này đặc biệt có lợi cho các tác nhân mã hóa, vì nó cho phép họ tập trung vào nhiệm vụ chính

> của mình trong khi dựa vào Code Finder để xử lý quá trình tìm kiếm. Tổng thể, Code Finder nhằm mục đích đơn giản hóa và tăng tốc quá trình tìm kiếm mã cho các tác nhân mã hóa, khiến nó trở thành một công cụ hiệu quả và có giá trị trong quy trình làm việc của họ.

TIPS & TRICKS CHO DEV

Tối ưu hóa RAG

Vấn đề: Dữ liệu không đầy đủ cho mô hình AI.

Cách làm: Sử dụng RAG để thu thập và tăng cường dữ liệu, ví dụ với prompt "Tóm tắt bài viết về AI".

Đánh giá: Hiệu quả cho dữ liệu hạn chế, nên dùng khi cần thông tin cụ thể.

Tìm kiếm семантис

Vấn đề: Tìm kiếm thông tin không chính xác.

Cách làm: Sử dụng semantic search với embeddings để tìm kiếm ý nghĩa, ví dụ lệnh CLI "pip install sentence-transformers".

Đánh giá: Hiệu quả cho tìm kiếm phức tạp, nên dùng khi cần thông tin liên quan.

Tăng cường embeddings

Vấn đề: Dữ liệu không đủ đa dạng cho mô hình AI.

Cách làm: Sử dụng kỹ thuật tăng cường embeddings để đa dạng hóa dữ liệu, ví dụ với lệnh "python -m transformers.Trainer".

Đánh giá: Hiệu quả cho dữ liệu hạn chế, nên dùng khi cần cải thiện độ chính xác.

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 mô hình ngôn ngữ lớn (LLM). Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý, từ đó cải thiện trải nghiệm người dùng. Việc tối ưu hóa LLM cũng giúp giảm thiểu tác động môi trường của ứng dụng.
3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning để điều chỉnh mô hình cho phù hợp với nhiệm vụ cụ thể, hoặc sử dụng LoRA (Low-Rank Adaptation) để giảm kích thước mô hình mà không ảnh hưởng đến hiệu suất.
4. Tip hoặc bước tiếp theo: Để bắt đầu tối ưu hóa LLM, hãy bắt đầu bằng việc phân tích chi phí tính toán và hiệu năng của mô hình hiện tại, sau đó thử nghiệm các kỹ thuật như fine-tuning, LoRA hoặc quantization để tìm ra giải pháp tối ưu cho ứng dụng của mình.

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

Nguồn: Google News · Groq AI