

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Cái khó ló cái khôn.”

— Tục ngữ Việt Nam

Áp lực và khó khăn buộc ta phải sáng tạo và tìm ra giải pháp mới — nghịch cảnh là bà đỡ của sự đổi mới.

TIN TỨC NỔI BẬT

1

Xây dựng agent phân tích tài chính thông minh với LangGraph và Strands Agents

Build an intelligent financial analysis agent with LangGraph and Strands Agents

Amazon Web Services (AWS) [Đọc bài viết →](#)

Amazon Web Services (AWS) đã giới thiệu một giải pháp để xây dựng một agent phân tích tài chính thông minh sử dụng LangGraph và Strands Agents. Cách tiếp cận đổi mới này cho phép người dùng tạo ra một công cụ phân tích tài chính tinh vi có khả năng xử lý và phân tích lượng lớn dữ liệu tài chính. LangGraph là một thư viện xử lý ngôn ngữ tự nhiên (NLP) cho phép người dùng trích xuất thông tin liên quan từ các báo cáo, bảng kê và tài liệu tài chính khác. Ngược lại, Strands Agents là một nền tảng cho phép tạo ra các intelligent agent có thể thực hiện các tác vụ phức tạp, chẳng hạn như phân tích tài chính. Bằng cách kết hợp LangGraph và Strands Agents, người dùng có thể xây dựng một financial analysis agent có khả năng tự động trích xuất các chỉ số tài chính quan trọng, xác định xu hướng và mô hình, cũng như cung cấp các insight có thể hành động. Intelligent agent này có thể được tích hợp với nhiều hệ thống và công cụ tài chính khác nhau, giúp người dùng đưa ra các quyết định đầu tư sáng suốt hơn. Giải pháp này được thiết kế để giúp các chuyên gia tài chính, nhà đầu tư và tổ chức hợp lý hóa quy trình phân tích tài chính, giảm thiểu công sức thủ công và cải thiện việc ra quyết định.

2

LangChain ra mắt công cụ xây dựng agent mã nguồn mở "dễ tiếp cận"

LangChain đã giới thiệu một agent builder mã nguồn mở được thiết kế để dễ tiếp cận hơn với các developer. Công cụ mới này nhằm mục đích đơn giản hóa quy trình tạo và triển khai các AI-powered agent.

3

Các model open-weight tốt nhất của Trung Quốc — và đối thủ mạnh nhất của Mỹ

The best Chinese open-weight models — and the strongest US rivals

understandingai.org

[Đọc bài viết →](#)

Bài viết so sánh các model open-weight hàng đầu của Trung Quốc với các đối thủ mạnh nhất của Mỹ. Các model open-weight ở đây đề cập đến các phương tiện điện không có giới hạn trọng lượng, cho phép nhiều tùy chọn tùy chỉnh. Các model của Trung Quốc được nêu bật trong bài viết nằm trong số tốt nhất cùng phân khúc, mang lại hiệu suất, phạm vi hoạt động và tính năng ấn tượng. Bài viết không cung cấp chi tiết cụ thể về các model của Trung Quốc, nhưng gợi ý rằng chúng là những đối thủ mạnh trên thị trường. Ngược lại, các đối thủ của Mỹ cũng thuộc hàng top, mang lại hiệu suất và khả năng vượt trội. Bài viết ngụ ý rằng các model của Trung Quốc đang cạnh tranh gay gắt với các đối thủ Mỹ, thậm chí một số còn vượt trội hơn ở một số khía cạnh. Tuy nhiên, bài viết không đưa ra so sánh hoặc xếp hạng trực tiếp các model. Nhìn chung, bài viết cung cấp cái nhìn tổng quan về các model open-weight của Trung Quốc và các đối thủ Mỹ, làm nổi bật sự cạnh tranh ngày càng tăng trên thị trường xe điện.

4

Thủ thư tổ chức workshop 'Tránh xa AI' gây sốt cho những người chán ngấy Big Tech

Librarians are hosting viral 'Avoiding AI' workshops for people who are fed up with Big Tech

TechCrunch AI

[Đọc bài viết →](#)

Các thủ thư trên khắp nước Mỹ đang tổ chức các workshop để giúp mọi người tránh sự hiện diện ngày càng tăng của Trí tuệ nhân tạo (AI) trong cuộc sống hàng ngày. Các workshop "Tránh xa AI", lấy cảm hứng từ một thủ thư ở Maine, cung cấp hướng dẫn từng bước về cách tắt các tính năng AI trên các nền tảng và thiết bị công nghệ phổ biến. Được dẫn dắt bởi các thủ thư như Charlie Bailey ở Philadelphia và

Hannah Cyrus ở Maine, các workshop này nhằm mục đích nâng cao kiến thức kỹ thuật số và trao quyền cho các cá nhân giành lại quyền kiểm soát việc sử dụng các công cụ AI của họ. Các workshop đã thu hút sự chú ý đáng kể, với hơn 2.000 lượt thích và 220 lượt chia sẻ trên Instagram cho sự kiện của Bailey và 70 người tham dự hai workshop đầu tiên của Cyrus, bao gồm cả một buổi livestream. Các buổi học thúc đẩy ý thức cộng đồng giữa những người tham gia, những người chia sẻ các mẹo và kinh nghiệm của riêng họ với AI. Bằng cách cung cấp các workshop này, các thủ thư đang giải quyết mối lo ngại ngày càng tăng về việc áp dụng các công cụ AI một cách cưỡng bức và cung cấp một nền tảng để mọi người đưa ra lựa chọn sáng suốt về cuộc sống số của họ.

5

Ra mắt tính năng Sức khỏe trên ChatGPT

Launching Health in ChatGPT

OpenAI Blog

[Đọc bài viết →](#)

ChatGPT đã mở rộng các tính năng sức khỏe của mình để cung cấp cho người dùng những insight cá nhân hóa hơn và hiểu biết sâu sắc hơn về sức khỏe của họ. Người dùng đủ điều kiện tại Hoa Kỳ giờ đây có thể kết nối an toàn hồ sơ y tế và dữ liệu Apple Health của họ với nền tảng. Việc tích hợp này nhằm mục đích cung cấp cho người dùng cái nhìn toàn diện hơn về sức khỏe của họ bằng cách kết hợp lịch sử y tế với dữ liệu Apple Health. Chức năng bổ sung này dự kiến sẽ giúp người dùng hiểu rõ hơn về sức khỏe tổng thể của mình và đưa ra các quyết định sáng suốt hơn về tình trạng sức khỏe của họ. Việc kết nối an toàn hồ sơ y tế và dữ liệu Apple Health là một khía cạnh quan trọng của tính năng này, đảm bảo rằng thông tin sức khỏe nhạy cảm của người dùng được bảo vệ.

6

Vượt xa RAG thụ động: Cách triển khai tái cấu trúc bộ nhớ chủ động cho các AI agent

Moving beyond passive RAG: How to implement active memory reconstruction for AI agents

BD Tech Talks

[Đọc bài viết →](#)

Các nhà nghiên cứu tại Đại học Quốc gia Singapore đã phát triển một framework mới gọi là MRAgent để cải thiện hiệu suất của các tác nhân AI trong các nhiệm vụ trò chuyện. Phương pháp "retrieve-then-reason" hiện tại, nơi các tác nhân AI thu thập thông tin từ bộ nhớ và

sau đó suy luận trên đó, bị giới hạn bởi bản chất tĩnh và đường ống thu hồi bộ nhớ cứng nhắc. MRAgent giải quyết vấn đề này bằng cách giới thiệu một cơ chế tái tạo bộ nhớ chủ động tích hợp tái tạo bộ nhớ vào quá trình suy luận của AI. Điều này cho phép tác nhân dần dần phát triển bộ nhớ của nó dựa trên bằng chứng và kinh nghiệm tích lũy. Các thí nghiệm đã chỉ ra rằng MRAgent đạt được cải thiện hiệu suất lên đến 23% trên các nhiệm vụ suy luận phức tạp, đồng thời giảm tiêu thụ token prompt và chi phí thời gian chạy. Framework này được thiết kế để cho phép các tác nhân AI mở rộng để sử dụng các trường hợp có tương tác dài hạn và cung cấp các phản hồi đáng tin cậy và nhận thức ngữ cảnh hơn. Bằng cách áp dụng một quá trình tái tạo chủ động và liên kết, MRAgent nhằm mục đích vượt qua các hạn chế của các đường ống thu hồi thụ động hiện tại và cho phép các tác nhân AI suy luận trên lịch sử trò chuyện dài hạn một cách hiệu quả hơn.

7

Midjourney mua lại ứng dụng chiêm tinh Co-Star

Midjourney bought the astrology app Co-Star

The Verge AI

[Đọc bài viết →](#)

Midjourney, một công ty khởi nghiệp AI nổi tiếng với khả năng tạo hình ảnh, đã mua lại ứng dụng chiêm tinh Co-Star. Các điều khoản của thỏa thuận, được báo cáo đã kết thúc vào mùa xuân, không được tiết lộ. Người sáng lập Co-Star, Banu Guler, sẽ vẫn phụ trách ứng dụng và cũng đảm nhận vai trò giám đốc thiết kế của Midjourney. Guler và đội ngũ của cô sẽ giúp Midjourney xây dựng ứng dụng đầu tiên của mình, bao gồm cả ứng dụng tạo hình ảnh. Hiện tại, Midjourney cung cấp các model AI của mình trên web và thông qua Discord. Co-Star là một ứng dụng miễn phí cung cấp horoscope hàng ngày và kiểm tra tính tương thích, kết hợp giữa cái nhìn sâu sắc của con người, dữ liệu của NASA và AI để cung cấp lời khuyên cá nhân hóa cho người dùng. Với việc mua lại này, Midjourney đang mở rộng vào lĩnh vực chiêm tinh, đánh dấu một hướng đi mới cho công ty.

8

Canary tokens và kỹ thuật số

Canary tokens and digital tripwires

Changelog

[Đọc bài viết →](#)

Người sáng lập Thinkst, Haroon Meer, thảo luận về các sản phẩm của công ty, bao gồm Canary và Canarytokens, là những bẫy và dây cò

được thiết kế để phát hiện và ngăn chặn các kẻ tấn công mạng. Meer nhấn mạnh hiệu quả của Canarytokens, đặc biệt là token khóa API AWS mà các kẻ tấn công thường cố gắng khai thác. Ông cũng giới thiệu Breadcrumbs, một tính năng mới giúp hướng dẫn những kẻ xâm nhập đến các canary, làm cho việc theo dõi và phân tích các cuộc tấn công trở nên dễ dàng hơn. Ngoài ra, Meer trình diễn trực tiếp việc biến một thiết bị Canary phần cứng thành một thiết bị lưu trữ mạng Synology chỉ trong một cú nhấp chuột. Sự thành công của Thinkst cũng được nhấn mạnh, với doanh thu hàng năm tái diễn (ARR) đạt 22,5 triệu đô la mà không cần bán hàng trực tiếp và không tăng giá trong mười năm.

TIPS & TRICKS CHO DEV

Cài đặt Ollama

Vấn đề: Không biết cách cài đặt Ollama trên máy cá nhân.

Cách làm: Sử dụng lệnh `pip install ollama` để cài đặt, sau đó chạy `ollama --help` để xem các tùy chọn. Ví dụ: `ollama --model lmstudio/bloom-350m`.

Đánh giá: Hiệu quả cao, giúp chạy LLM trên máy cá nhân.

Chạy LM Studio

Vấn đề: Không biết cách chạy LM Studio trên máy cá nhân.

Cách làm: Sử dụng lệnh `lm-studio --model bloom-350m` để chạy, sau đó nhập prompt như "Tôi cần viết một bài về chủ đề AI".

Đánh giá: Hiệu quả cao, giúp tạo ra nội dung chất lượng.

Tối ưu hóa tài nguyên

Vấn đề: LLM tiêu tốn nhiều tài nguyên máy tính.

Cách làm: Sử dụng lệnh `ollama --gpu` để sử dụng GPU, hoặc `lm-studio --cpu` để sử dụng CPU.

Đánh giá: Giúp tiết kiệm tài nguyên, chạy mượt trên máy yếu.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Để phát triển ứng dụng AI hiệu quả, việc tối ưu chi phí và hiệu năng của 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 kỹ thuật như fine-tuning và LoRA có thể giúp đạt được mục tiêu này.

3. Ví dụ, sử dụng kỹ thuật LoRA (Low-Rank Adaptation) để điều chỉnh mô hình LLM cho một nhiệm vụ cụ thể, giúp giảm số lượng tham số cần huấn luyện và tăng

tốc độ huấn luyện.

4. Tip hoặc bước tiếp theo: Áp dụng kỹ thuật fine-tuning và LoRA để tối ưu hóa mô hình LLM cho ứng dụng của bạn, và đánh giá hiệu suất của mô hình sau khi áp dụng các kỹ thuật này.

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

Nguồn: Google News · Groq AI