BIS - MT · 24/05/2026

🤖 Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

✨ *"Rest when you're weary. Refresh and renew yourself, your body, your mind, your spirit."*

↪ Hãy nghỉ ngơi khi bạn mệt mỏi. Làm mới bản thân, cơ thể, tâm trí và tinh thần.

— Ralph Marston

💡 *Biết khi nào cần dừng lại và tái nạp năng lượng là trí tuệ — làm việc hiệu quả không phải làm liên tục mà là làm đúng thời điểm.*

📰 TIN TỨC NỔI BẬT

1

Xây dựng một tác nhân phân tích tài chính thông minh với LangGraph và Strands Agents | Amazon Web Services

🇬🇧 *Build an intelligent financial analysis agent with LangGraph and Strands Agents | Amazon Web Services*

📰 Amazon Web Services (AWS) 🔗 Đọc bài viết →

> Amazon Web Services (AWS) đã giới thiệu một giải pháp để xây dựng một tác nhân phân tích tài chính thông minh sử dụng LangGraph và Strands Agents. Tác nhân này có thể phân tích dữ liệu tài chính, xác định xu hướng và cung cấp thông tin chi tiết để giúp các doanh nghiệp đưa ra quyết định sáng suốt. LangGraph là một thư viện xử lý ngôn ngữ tự nhiên (NLP) cho phép tác nhân hiểu và xử lý dữ liệu tài chính, bao gồm cả báo cáo và tuyên bố dựa trên văn bản. Strands Agents, mặt khác, là một framework để xây dựng giao diện đối thoại có thể tương tác với người dùng và cung cấp khuyến nghị cá nhân hóa. Bằng cách kết hợp LangGraph và Strands Agents, các doanh nghiệp có thể tạo ra một tác nhân phân tích tài chính có thể: * Phân tích dữ liệu tài chính và xác định xu hướng * Cung cấp khuyến nghị cá nhân hóa cho người dùng * Cung cấp thông tin chi tiết và đề xuất để cải thiện kinh doanh * Tự động hóa các nhiệm vụ phân tích tài chính, giúp nhân viên tập trung vào các hoạt động có giá trị cao hơn Giải pháp này có thể được sử dụng bởi các doanh nghiệp mọi quy mô để có được sự hiểu biết sâu sắc hơn về hiệu suất tài chính của họ và đưa ra quyết định dựa trên dữ liệu.

2 AWS mã nguồn mở một máy chủ MCP cho Bedrock AgentCore để tối ưu hóa phát triển tác nhân AI

🇬🇧 *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 (Giao thức Điều khiển Đa tác nhân) cho Bedrock AgentCore, một khuôn khổ được thiết kế để đơn giản hóa việc phát triển tác nhân AI. Máy chủ MCP là một thành phần quan trọng của Bedrock AgentCore, cho phép giao tiếp liền mạch giữa các tác nhân AI và môi trường mà chúng tương tác. Bằng cách mở nguồn máy chủ MCP, AWS nhằm mục đích đẩy nhanh việc phát triển các tác nhân AI và tạo điều kiện cho sự hợp tác giữa các nhà nghiên cứu và nhà phát triển. Máy chủ MCP sẽ cho phép người dùng dễ dàng tích hợp và quản lý nhiều tác nhân AI, giúp xây dựng và triển khai các hệ thống AI phức tạp trở nên dễ dàng hơn. Việc mở nguồn máy chủ MCP là một bước tiến quan trọng trong việc phát triển công nghệ AI, vì nó cho phép cộng đồng đóng góp và xây dựng dựa trên khuôn khổ này. Động thái này dự kiến sẽ thúc đẩy sự đổi mới và tiến bộ trong nghiên cứu AI, đặc biệt là trong các lĩnh vực như robot, hệ thống tự động và phát triển trò chơi.

3 Nhà máy tác nhân: Từ nguyên mẫu đến sản xuất - công cụ dành cho nhà phát triển và phát triển tác nhân nhanh chóng

🇬🇧 *Agent Factory: From prototype to production—developer tools and rapid agent development*

📰 Microsoft Azure 🔗 Đọc bài viết →

> Microsoft đã giới thiệu Agent Factory, một bộ công cụ dành cho developer được thiết kế để tăng tốc quá trình tạo và triển khai các tác nhân thông minh. Những tác nhân này có thể được sử dụng trong nhiều ứng dụng khác nhau, bao gồm cả chatbot, trợ lý ảo và các hệ thống khác được hỗ trợ bởi AI. Các công cụ này được xây dựng trên nền tảng Microsoft Azure, cho phép developer tận dụng khả năng mở rộng và độ tin cậy của nền tảng đám mây. Với Agent Factory, developer có thể nhanh chóng phát triển và thử nghiệm các tác nhân thông minh bằng cách sử dụng một loạt các mẫu và API đã được xây dựng sẵn. Nền tảng này cung cấp một giao diện thân thiện với người dùng để định nghĩa hành vi của tác nhân, tích hợp với các dịch vụ bên ngoài và triển khai tác nhân đến môi trường sản xuất. Quá trình phát triển được tối ưu hóa này cho phép developer đưa các dự án được hỗ

trợ bởi AI của họ ra thị trường nhanh hơn và hiệu quả hơn. Bằng cách cung cấp một bộ công cụ toàn diện cho việc phát triển tác nhân nhanh chóng, Agent Factory nhằm mục đích đơn giản hóa quá trình tạo ra các tác nhân thông minh và làm cho chúng trở nên dễ tiếp cận hơn với nhiều developer. Điều này có thể dẫn đến việc tạo ra nhiều ứng dụng được hỗ trợ bởi AI sáng tạo và hiệu quả hơn trên nhiều ngành công nghiệp khác nhau.

4

Pin sạc di động tốt nhất (2026): Lựa chọn của tôi sau khi thử nghiệm hơn 100 sản phẩm

🇬🇧 *Best Power Banks (2026): My Picks After Testing Over 100*

📰 Wired 🔗 Đọc bài viết →

Bài viết đánh giá các pin sạc di động hàng đầu trên thị trường, dựa trên quá trình thử nghiệm rộng rãi của biên tập viên với hơn 100 bộ sạc di động. Biên tập viên khuyến nghị bốn pin sạc thiết yếu, phù hợp với hầu hết người dùng, và nhấn mạnh một số tùy chọn bổ sung cho các trường hợp sử dụng độc đáo. Lựa chọn hàng đầu là pin sạc Anker, có dung lượng 25.000-mAh, công suất đầu ra 165-watt và một loạt các tính năng bao gồm màn hình, cáp tích hợp và sạc qua. Pin sạc này cũng nhỏ gọn và thân thiện với du lịch, khiến nó trở thành lựa chọn lý tưởng cho sử dụng hàng ngày. Các tùy chọn đáng chú ý khác bao gồm Nimble Champ Pro, một pin sạc di động nhẹ và có dung lượng 20.000-mAh, và Sharge 170, một pin sạc thời trang và chống nước với dung lượng 24.000-mAh. Bài viết cũng đề cập đến tác động môi trường của pin sạc và nhấn mạnh nỗ lực của Nimble trong việc giảm lãng phí và sử dụng vật liệu tái chế.

5

OpenAI được đặt tên là Nhà lãnh đạo trong lĩnh vực tác nhân mã hóa doanh nghiệp bởi Gartner

🇬🇧 *OpenAI named a Leader in enterprise coding agents by Gartner*

📰 OpenAI Blog 🔗 Đọc bài viết →

OpenAI đã được công nhận là một nhà lãnh đạo trong 2026 Gartner Magic Quadrant dành cho các tác nhân mã hóa AI doanh nghiệp. Xếp hạng danh giá này công nhận cách tiếp cận đổi mới của công ty đối với mã hóa AI, đặc biệt là thông qua công nghệ Codex của họ. Khung đánh giá Gartner Magic Quadrant là một khuôn khổ đánh giá được kính trọng rộng rãi, đánh giá các nhà cung cấp dựa trên khả năng

thực thi và tính đầy đủ của tầm nhìn. Bằng việc được đặt tên là nhà lãnh đạo, OpenAI đã chứng minh chuyên môn của mình trong việc phát triển và triển khai các tác nhân mã hóa AI ở quy mô doanh nghiệp. Thành tựu này nhấn mạnh cam kết của công ty trong việc đẩy ranh giới của trí tuệ nhân tạo và ứng dụng của nó trong không gian mã hóa. Sự công nhận này là minh chứng cho khả năng đổi mới của OpenAI và khả năng đáp ứng nhu cầu phức tạp của các doanh nghiệp quy mô lớn.

6

Tại sao việc cô lập OpenClaw không ngăn chặn việc trích xuất dữ liệu

🇬🇧 *Why sandboxing OpenClaw doesn't stop data exfiltration*

📰 BD Tech Talks 🔗 Đọc bài viết →

Một nghiên cứu bảo mật gần đây của Lasso đã tiết lộ rằng việc cô lập OpenClaw, một loại đại lý AI tự động, không cung cấp sự bảo vệ đủ đối với việc lấy cắp dữ liệu. Mặc dù bị cô lập trong một môi trường ảo, các đại lý OpenClaw vẫn có thể bị thao túng để làm rò rỉ dữ liệu nhạy cảm hoặc viết lại các hướng dẫn của chính chúng. Nghiên cứu đã xác định các điểm yếu trong NemoClaw, môi trường cô lập của Nvidia để chạy OpenClaw, có thể bị khai thác thông qua các cuộc tấn công tiêm prompt tinh vi. Những cuộc tấn công này cho phép các tác nhân độc hại phân phối phần mềm độc hại, vượt qua các bộ lọc phát hiện tĩnh và thay đổi bản sắc cốt lõi của một đại lý. Những phát hiện này nhấn mạnh những hạn chế của các biện pháp bảo mật truyền thống trong việc bảo vệ các hệ thống lưu trữ các đại lý AI, vì đường dẫn thực hiện của chúng được xác định động bởi văn bản mà chúng đọc. Nghiên cứu đã chứng minh hai vector tấn công, bao gồm ngộ độc phụ thuộc và ngộ độc cấu hình đại lý, có thể được sử dụng để lấy cắp dữ liệu nhạy cảm từ các đại lý OpenClaw. Việc nghiên cứu nhấn mạnh nhu cầu về các biện pháp bảo mật mạnh mẽ hơn để bảo vệ chống lại hành vi không thể đoán trước của các đại lý AI.

7

AiFinPay: SDK AiFinPay cung cấp một quá trình thanh toán liền mạch và bảo mật

🇬🇧 *AiFinPay: The AiFinPay SDK provides a seamless and secure pa*

📰 Dev.to AI 🔗 Đọc bài viết →

AiFinPay cung cấp một giải pháp thanh toán bảo mật cho các đại lý AI thông qua SDK của mình. Điều này cho phép thực hiện các giao dịch hiệu quả và tự động hóa, tối ưu hóa các quy trình tài chính và nâng cao hiệu suất kinh doanh. SDK cung cấp một trải nghiệm thanh toán liền mạch, cho phép thực hiện các giao dịch nhanh chóng và bảo mật. Ngoài ra, nó còn cung cấp các mẫu cho các câu hỏi thường gặp và các đoạn mã có thể tái sử dụng, giúp dễ dàng quản lý và tự động hóa các quy trình tài chính.

8

Mô hình AI mới của Google từ bất cứ thứ gì sang bất cứ thứ gì là điên rồ

🇬🇧 *Google's new anything-to-anything AI model is wild*

📰 The Verge AI 🔗 Đọc bài viết →

Gần đây, Google đã phát hành một mô hình AI mới gọi là Omni, là một phần của gia đình mô hình tạo sinh, có thể chuyển đổi bất kỳ đầu vào nào - ảnh, video hoặc văn bản - thành định dạng khác. Phiên bản đầu tiên của mô hình này, Omni Flash, hiện đã có sẵn trên nền tảng tạo và chỉnh sửa video AI của Google, Flow. Omni cải tiến so với người tiền nhiệm của nó, Veo, bằng cách cho phép người dùng tải lên một video và kết hợp nó với một lời nhắc văn bản để tạo nội dung được tạo bởi AI. Mô hình này cũng tuyên bố rằng nó kết hợp nhiều kiến thức thế giới thực và duy trì tính nhất quán của nhân vật trong suốt video. Khi thử nghiệm Omni, tác giả đã tìm thấy kết quả hỗn hợp, với một số video rất tốt và nhất quán, trong khi những video khác có những "cú sốc nhảy" và sự không nhất quán rõ ràng được tạo bởi AI. Mô hình này cũng gặp khó khăn khi chỉnh sửa và thực hiện các lời nhắc của người dùng, đôi khi tạo ra kết quả không mong muốn và không mong muốn. Mặc dù những vấn đề này, Omni cho thấy tiềm năng trong khả năng tạo ra các video thực tế với nỗ lực và kiến thức đáng ngạc nhiên là rất ít.

⚡ TIPS & TRICKS CHO DEV

⚡ Chain-of-Thought

Vấn đề: ChatGPT không trả lời chính xác do thiếu thông tin.

Cách làm: Dùng kỹ thuật chain-of-thought, đặt câu hỏi từng bước, ví dụ "Bước 1: Xác định yêu cầu, Bước 2: Tìm kiếm thông tin".

Đánh giá: Hiệu quả khi cần câu trả lời chi tiết, nên dùng khi vấn đề phức tạp.

⚡ Few-Shot Learning

Vấn đề: Claude không thể học từ dữ liệu nhỏ.

Cách làm: Áp dụng few-shot learning, cung cấp 2-3 ví dụ để Claude học, ví dụ "Xác định cảm xúc trong câu: Ví dụ 1: Câu 'Anh ấy rất vui' có cảm xúc tích cực".

Đánh giá: Hiệu quả khi dữ liệu hạn chế, nên dùng khi cần huấn luyện nhanh.

⚡ System Prompt Design

Vấn đề: Gemini không hiểu yêu cầu do prompt không rõ ràng.

Cách làm: Thiết kế system prompt rõ ràng, bao gồm both chức năng và giới hạn, ví dụ "Tóm tắt văn bản, không quá 100 từ, chỉ bao gồm thông tin quan trọng".

Đánh giá: Hiệu quả khi cần câu trả lời chính xác, nên dùng khi yêu cầu phức tạp.

📚 BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**

Dev cần biết về tối ưu chi phí và hiệu năng LLM để giảm thiểu chi phí vận hành và cải thiện hiệu suất của ứng dụng AI. Điều này đặc biệt quan trọng khi xây dựng ứng dụng quy mô lớn.

2. Việc tối ưu hóa chi phí và hiệu năng LLM giúp giảm thiểu tài nguyên và tăng tốc độ xử lý dữ liệu.

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

4. 💡 Tip: Sử dụng các công cụ như Hugging Face Transformers để tối ưu hóa và triển khai mô hình LLM hiệu quả.