BIS - MT · 02/06/2026

🤖 Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

✨ *"I'm not afraid of storms, for I'm learning how to sail my ship."*

↪ Tôi không sợ bão tố, vì tôi đang học cách lái con thuyền của mình.

— Louisa May Alcott

💡 *Khó khăn và thử thách là bài học để trưởng thành — thay vì tránh né, hãy xem mỗi 'cơn bão' như cơ hội để tôi luyện bản lĩnh.*

📰 TIN TỨC NỔI BẬT

1

Nhà máy Agent: Từ nguyên mẫu đến sản xuất - công cụ dành cho nhà phát triển và phát triển agent 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 công cụ dành cho nhà phát triển được thiết kế để tối ưu hóa quá trình tạo và triển khai các tác nhân thông minh. Agent Factory cho phép các nhà phát triển nhanh chóng tạo mẫu và thử nghiệm các tác nhân thông minh, sau đó có thể dễ dàng triển khai vào môi trường sản xuất. Công cụ này được xây dựng trên nền tảng Microsoft Azure, cho phép các nhà phát triển tận dụng khả năng mở rộng và độ tin cậy của nền tảng đám mây. Với Agent Factory, các nhà phát triển có thể tạo, đào tạo và triển khai các tác nhân thông minh bằng cách sử dụng một loạt các mẫu và công cụ đã được xây dựng sẵn. Nền tảng này cung cấp một giao diện người dùng thân thiện để xây dựng và thử nghiệm các tác nhân, giúp các nhà phát triển tập trung vào logic và hành vi của các tác nhân của họ thay vì cơ sở hạ tầng cơ bản. Bằng cách cung cấp một môi trường phát triển nhanh chóng, Agent Factory nhằm mục đích đẩy nhanh việc áp dụng các tác nhân thông minh trong các ngành công nghiệp khác nhau, bao gồm dịch vụ khách hàng, quản lý chuỗi cung ứng và nhiều hơn nữa. Công cụ này được thiết kế để giúp các nhà phát triển xây dựng các tác nhân thông minh hiệu quả, có thể mở rộng và bảo mật hơn, cuối cùng thúc đẩy giá trị kinh doanh và đổi mới.

2 So sánh 7 mô hình ngôn ngữ lớn hàng đầu LLM/Hệ thống cho việc lập trình năm 2025

🇬🇧 *Comparing the Top 7 Large Language Models LLMs/Systems for Coding in 2025*

📰 MarkTechPost 🔗 Đọc bài viết →

> Vào năm 2025, phong cảnh lập trình đã chứng kiến những tiến bộ đáng kể với sự xuất hiện của các Large Language Models (LLMs) hàng đầu. Một so sánh gần đây đã nhấn mạnh 7 LLM hàng đầu cho lập trình,□□ khả năng và điểm mạnh của chúng. Những model này bao gồm LLaMA, PaLM 2 và BLOOM, đã thể hiện hiệu suất vượt trội trong các nhiệm vụ lập trình. LLaMA, được phát triển bởi Meta, đã cho thấy kết quả ấn tượng trong hoàn thành mã và gỡ lỗi. PaLM 2, mặt khác, đã xuất sắc trong việc tạo mã và đã được sử dụng trong nhiều ứng dụng. BLOOM, được phát triển bởi Anthropic, cũng đã thể hiện hiệu suất mạnh mẽ trong các nhiệm vụ lập trình. Các model đáng chú ý khác bao gồm Llama 2, đã cải thiện khả năng của người tiền nhiệm, và Chinchilla, đã thể hiện hiệu suất đáng kinh ngạc trong các nhiệm vụ lập trình. Ngoài ra, so sánh cũng nhấn mạnh điểm mạnh của OPT-175B và Codex, đã được sử dụng trong nhiều ứng dụng lập trình. Những model này có tiềm năng cách mạng hóa quá trình lập trình, mang lại kết quả nhanh hơn và chính xác hơn.

3 Claude vs. ChatGPT Thống kê 2026: Số liệu đối đầu đằng sau trận chiến AI

🇬🇧 *Claude vs. ChatGPT Statistics 2026: Head-to-Head Numbers Behind the AI Battle*

📰 SQ Magazine 🔗 Đọc bài viết →

> Một so sánh gần đây giữa hai mô hình AI nổi bật, Claude và ChatGPT, đã làm sáng tỏ các thống kê hiệu suất của chúng vào năm 2026. Theo dữ liệu, Claude có thời gian phản hồi là 100 mili giây, trong khi ChatGPT mất khoảng 200 mili giây để phản hồi. Về độ chính xác, Claude đạt được tỷ lệ chính xác là 92%, vượt qua tỷ lệ 88% của ChatGPT. Các mô hình cũng được so sánh về khả năng hiểu và phản hồi các truy vấn phức tạp. Claude đã chứng minh tỷ lệ thành công 95% trong việc xử lý các câu hỏi phức tạp, trong khi ChatGPT quản lý được 85%. Ngoài ra, cơ sở kiến thức của Claude được báo cáo là lớn hơn 15% so với cơ sở kiến thức của ChatGPT, cho thấy phạm vi chủ đề và thông tin rộng lớn hơn. So sánh cũng làm nổi bật sự khác biệt trong phong cách tương tác của các mô hình. Claude được mô tả là trực tiếp và đi thẳng vào vấn đề trong các phản hồi của nó, trong khi ChatGPT

có xu hướng cung cấp các câu trả lời chi tiết và phức tạp hơn. Những thống kê này cung cấp thông tin chi tiết quý giá về hiệu suất và khả năng của hai mô hình AI này, làm nổi bật điểm mạnh và điểm yếu của chúng trong các lĩnh vực khác nhau.

4

Vẫn là nhà phát triển. Chỉ bên ngoài. Bộ sưu tập cửa hàng GitHub mới nhất của chúng tôi đã có mặt.

🇬🇧 *Still a developer. Just outside. Our latest GitHub Shop collection is here.*

📰 GitHub Blog 🔗 Đọc bài viết →

GitHub đã phát hành một bộ sưu tập hàng hóa mới, được gọi là ESC (Escape the Confines), lấy cảm hứng từ ý tưởng bước ra khỏi bàn làm việc để kích thích sự sáng tạo và giải quyết vấn đề. Bộ sưu tập này bao gồm nhiều mặt hàng như mũ, tất, áo t-shirt và dép hồ bơi, tất cả đều có các nhân vật biểu tượng của GitHub. Bộ sưu tập ESC được thiết kế để giúp các developer nghỉ ngơi khỏi công việc và tìm kiếm cảm hứng trong môi trường mới. Bộ sưu tập này cũng trùng hợp với nỗ lực của GitHub trong việc□□ khả năng và lợi ích của trí tuệ nhân tạo (AI), học máy (machine learning), và DevSecOps, cũng như vị trí lãnh đạo của mình trong không gian nền tảng developer. Bộ sưu tập hiện đã có sẵn trong cửa hàng của GitHub, mang đến cơ hội cho các developer thể hiện sự sáng tạo và□□ tình yêu của họ dành cho nền tảng này.

5

MiniMax-M3 ra mắt, vượt qua GPT-5.5 và Gemini 3.1 Pro về hiệu suất chuẩn mực chính với chỉ 5-10% chi phí

🇬🇧 *MiniMax-M3 debuts, eclipsing GPT-5.5 and Gemini 3.1 Pro on key benchmark performance for just 5-10% of the cost*

📰 VentureBeat 🔗 Đọc bài viết →

MiniMax, một công ty khởi nghiệp AI của Trung Quốc, đã phát hành mô hình ngôn ngữ lớn M3 được mong đợi cao, đã vượt qua các mô hình độc quyền hàng đầu từ Google, OpenAI và Anthropic về hiệu suất chuẩn mực chính tại chi phí thấp hơn đáng kể. Mô hình này có sẵn với giá 0,3 đô la cho 1 triệu token đầu vào và 1,20 đô la cho 1 triệu token đầu ra trong tuần tới, và giá đầy đủ là 0,6 đô la / 2,40 đô la cho 1 triệu token đầu vào / đầu ra. Đây là một phần nhỏ của chi phí của các mô hình độc quyền hàng đầu, có thể có giá lên đến 8-20% giá của M3. Mô hình M3 được hỗ trợ bởi một kiến trúc mới khác biệt với mạng Transformer cổ điển, được gọi là Chú ý thưa thớt MiniMax (MSA), giúp giữ chi phí của mô hình thấp bằng cách giảm nhu cầu tính toán xuống

1/20 so với mô hình thế hệ trước. Mô hình này cũng có một hệ thống đa phương tiện bản địa, cho phép nó dịch các hình học trực quan phức tạp thành mã cấu trúc mà không mất tính trung thực ngữ cảnh. M3 đã đạt được kết quả ấn tượng trên các đánh giá tiêu chuẩn, bao gồm 59,0% trên SWE-Bench Pro, 66,0% trên Terminal Bench 2.1, 74,2% trên MCP Atlas và 83,5 trên BrowseComp. Tuy nhiên, nó tụt lại phía sau Claude Opus 4.8 của Anthropic trên một số chuẩn mực, đặc biệt là trong việc sửa đổi mã thuần túy và môi trường hệ thống tự động. MiniMax cũng đang phát hành mô hình M3 dưới giấy phép mã nguồn mở, cho phép tải xuống và tùy chỉnh toàn bộ doanh nghiệp miễn phí. Động thái này dự kiến sẽ làm cho mô hình trở nên hấp dẫn hơn cho sử dụng doanh nghiệp, vì nó sẽ cho phép các tổ chức chạy mô hình cục bộ trên phần cứng nội bộ, loại bỏ rủi ro rò rỉ dữ liệu liên quan đến API công khai.

6

Tự động hóa bảo mật phát triển: Từ SlackOps đến Triage SIEM lập trình (Phần 1/2)

🇬🇧 *Security Automation Evolved: From SlackOps to Programmatic SIEM Triage (Part 1/2)*

📰 Sourcegraph Blog 🔗 Đọc bài viết →

Đội ngũ an ninh của Sourcegraph đã trải qua một sự thay đổi đáng kể trong quy trình phản hồi sự cố của họ. Ban đầu, họ dựa vào bot phân loại trên Slack để quản lý các sự cố an ninh. Tuy nhiên, cách tiếp cận này có những hạn chế và cuối cùng đã được thay thế bằng một hệ thống chương trình tiên tiến hơn. Hệ thống mới này sử dụng phát hiện Security Information and Event Management (SIEM), cho phép đội ngũ tự động hóa quy trình phân loại. Điều này liên quan đến việc sử dụng các quy tắc tự động đóng dựa trên biểu thức để nhanh chóng xác định và giải quyết các sự cố an ninh. Bằng cách tận dụng công nghệ SIEM, đội ngũ an ninh có thể tối ưu hóa phản hồi của họ đối với các mối đe dọa tiềm năng, giảm thời gian và công sức cần thiết để điều tra và giải quyết các sự cố. Sự chuyển dịch này hướng tới phát hiện SIEM chương trình có thể đã cải thiện hiệu quả và hiệu quả của các hoạt động an ninh của Sourcegraph. Việc sử dụng các quy tắc tự động đóng dựa trên biểu thức cho phép đội ngũ tự động hóa các nhiệm vụ thường xuyên, giải phóng tài nguyên để tập trung vào các vấn đề an ninh phức tạp và ưu tiên cao hơn.

7

Quan điểm của chúng tôi về chính sách AI và vận động chính trị

🇬🇧 *Our views on AI policy and political advocacy*

📰 OpenAI Blog 🔗 Đọc bài viết →

> Công ty chúng tôi có quan điểm rõ ràng về chính sách trí tuệ nhân tạo (AI) và vận động chính trị. Chúng tôi ưu tiên tính minh bạch trong cách tiếp cận của mình, đảm bảo rằng quan điểm và hành động của chúng tôi là cởi mở và dễ hiểu đối với công chúng. Chúng tôi cũng ủng hộ việc quản lý AI một cách suy nghĩ, nhận ra tác động tiềm năng của nó đối với xã hội và nhu cầu xem xét cẩn thận. Hơn nữa, chúng tôi nhấn mạnh tầm quan trọng của sự an toàn của AI, thừa nhận các rủi ro tiềm ẩn liên quan đến công nghệ này. Điều đáng chú ý là chúng tôi không cho phép các nhóm chính trị bên ngoài nói thay mặt chúng tôi, duy trì sự độc lập và tự chủ trong các quy trình ra quyết định của mình. Cách tiếp cận này cho phép chúng tôi tham gia vào các cuộc thảo luận thông tin và có trách nhiệm về chính sách AI và vận động, đồng thời duy trì các giá trị của tính minh bạch và trách nhiệm. Bằng cách thực hiện một cách tiếp cận suy nghĩ và có nguyên tắc, chúng tôi nhằm mục đích đóng góp vào sự phát triển của AI theo cách mang lại lợi ích cho toàn xã hội.

8

Hàng chục gói Red Hat bị cài backdoor thông qua kênh NPM chính thức của nó

🇬🇧 *Dozens of Red Hat packages backdoored through its official NPM channel*

📰 Ars Technica 🔗 Đọc bài viết →

> Kênh NPM chính thức của Red Hat đã bị xâm phạm, cho phép một loại sâu độc hại lan truyền và đánh cắp thông tin đăng nhập nhạy cảm. Cuộc tấn công, bắt đầu từ thứ Hai, được thực hiện thông qua kênh @redhat-cloud-services, được các nhà phát triển tin cậy những người phụ thuộc vào dịch vụ đám mây Red Hat. Sâu độc hại, được đặt tên là Shai-Hulud, thực thi một payload bị che giấu trong quá trình cài đặt npm, thu thập thông tin đăng nhập nhạy cảm như bí mật hành động GitHub và thông tin đăng nhập dịch vụ đám mây. Malware sau đó lan truyền bằng cách xuất bản lại các gói bị nhiễm backdoor lên tài khoản của bên thứ ba. Hơn 30 gói đã bị ảnh hưởng và hầu hết đã bị gỡ xuống. Tuy nhiên, các tổ chức đã cài đặt một trong các gói bị ảnh hưởng được khuyến cáo nên coi hệ thống của mình là có khả năng bị xâm phạm. Cuộc tấn công này liên quan đến một cuộc tấn công chuỗi cung ứng trước đó đã nhiễm malware vào máy của một nhân viên, và Red Hat đã loại bỏ các gói độc hại. Công ty đã tuyên bố rằng cuộc

> điều tra của họ vẫn đang tiếp tục, nhưng cho đến nay, không có tác động nào đã được xác định đối với môi trường khách hàng hoặc đối tác hoặc hệ thống sản xuất của Red Hat.

⚡ TIPS & TRICKS CHO DEV

⚡ Tối ưu hóa LangGraph

Vấn đề: Xử lý task phức tạp với nhiều AI agents.
Cách làm: Sử dụng LangGraph để phối hợp nhiều AI agents, ví dụ: "Tạo một graph với các node là AI agents và edge là luồng dữ liệu giữa chúng".
Đánh giá: Hiệu quả cao khi xử lý task phức tạp, nhưng cần thiết kế graph phù hợp.

⚡ CrewAI cho tác vụ song song

Vấn đề: Xử lý nhiều task đồng thời với CrewAI.
Cách làm: Sử dụng CrewAI để tạo ra các agents song song, ví dụ: "Tạo một crew với các agents thực hiện task khác nhau".
Đánh giá: Tăng tốc độ xử lý task, nhưng cần quản lý tài nguyên hiệu quả.

⚡ AutoGen với phiData

Vấn đề: Tự động hóa tạo dữ liệu với AutoGen và phiData.
Cách làm: Sử dụng AutoGen để tạo dữ liệu và phiData để quản lý, ví dụ: "Tạo một pipeline với AutoGen và phiData".
Đánh giá: Tiết kiệm thời gian và tăng hiệu suất, nhưng cần kiểm tra dữ liệu tạo ra.

📚 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 áp dụng các kỹ thuật phù hợp. Điều này giúp giảm thiểu chi phí tính toán và bộ nhớ mà vẫn duy trì hiệu suất của mô hình.
3. Ví dụ, việc áp dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và tăng tốc độ luyện.
4. 💡 Tip hoặc bước tiếp theo: Hãy thử áp dụng kỹ thuật fine-tuning và LoRA vào mô hình LLM của bạn để tối ưu hóa hiệu năng và giảm chi phí, đồng thời nghiên cứu các phương pháp mới để tiếp tục cải thiện hiệu suất của mô hình.