BIS - MT · 23/05/2026

🤖 Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

✨ *"Almost everything will work again if you unplug it for a few minutes, including you."*

↪ Hầu hết mọi thứ sẽ hoạt động trở lại nếu bạn ngắt kết nối chúng vài phút, kể cả bạn.

— Anne Lamott

💡 *Nghỉ ngơi đúng lúc là kỹ năng quan trọng — tâm trí và cơ thể cần thời gian phục hồi để duy trì hiệu suất cao bền vững.*

📰 TIN TỨC NỔI BẬT

1

Claude Code so sánh với GitHub Copilot 2026: SWE-bench, Giá cả [Đã kiểm tra]

🇬🇧 *Claude Code vs GitHub Copilot 2026: SWE-bench, Pricing [Tested]*

📰 tech-insider.org 🔗 Đọc bài viết →

> Trong một so sánh gần đây, Claude Code và GitHub Copilot đã được kiểm tra về hiệu suất và giá cả của chúng. Cả hai công cụ đều là trợ lý lập trình AI được thiết kế để hỗ trợ các nhà phát triển phần mềm trong công việc của họ. Công cụ SWE-bench, một công cụ chuẩn hóa hiệu suất, đã được sử dụng để đánh giá hiệu suất của hai công cụ này. Theo kết quả, Claude Code đã vượt trội so với GitHub Copilot ở nhiều lĩnh vực, bao gồm hoàn thành mã, xem xét mã và tạo mã. Khả năng xử lý ngôn ngữ tự nhiên tiên tiến của Claude Code cho phép nó hiểu và phản hồi tốt hơn với đầu vào của nhà phát triển. Tuy nhiên, mô hình giá của GitHub Copilot, cung cấp một kế hoạch miễn phí với các tính năng hạn chế, có thể hấp dẫn hơn đối với một số nhà phát triển. So sánh giá cả cho thấy kế hoạch trả phí của GitHub Copilot đắt hơn so với kế hoạch của Claude Code, với chi phí hàng tháng là 49 đô la so với 29 đô la mỗi tháng của Claude Code. Kết quả gợi ý rằng Claude Code có thể là một lựa chọn tiết kiệm hơn cho các nhà phát triển cần hỗ trợ lập trình AI tiên tiến.

2 Các nhà nghiên cứu bảo mật phát tín hiệu cảnh báo về lỗ hổng trong mã được tạo bởi AI

🇬🇧 *Security Researchers Sound the Alarm on Vulnerabilities in AI-Generated Code*

📰 Infosecurity Magazine 🔗 Đọc bài viết →

> Các nhà nghiên cứu bảo mật đã bày tỏ mối quan ngại về các lỗ hổng trong mã được tạo ra bởi AI. Vấn đề này đã được nhấn mạnh tại Hội nghị RSA gần đây, nơi các chuyên gia đã thảo luận về các rủi ro tiềm ẩn liên quan đến mã được tạo ra bởi trí tuệ nhân tạo (AI) và các công cụ học máy (ML). Những mã này được tạo ra bởi AI đang ngày càng được sử dụng trong phát triển phần mềm, đặc biệt là trong các lĩnh vực như phân tích dữ liệu và tự động hóa. Tuy nhiên, các nhà nghiên cứu đã phát hiện ra rằng mã thường chứa các lỗ hổng, chẳng hạn như lỗi tiêm SQL và lỗi cross-site scripting (XSS), có thể bị khai thác bởi các kẻ tấn công. Các lỗ hổng này được cho là do thiếu sự giám sát và xem xét của con người trong quá trình tạo mã. Các công cụ AI và ML có thể tạo ra mã phức tạp và hiệu quả, nhưng chúng có thể không luôn tuân theo các phương pháp hay nhất hoặc hướng dẫn bảo mật. Do đó, các nhà nghiên cứu bảo mật đang kêu gọi các nhà phát triển thận trọng khi sử dụng mã được tạo ra bởi AI và xem xét kỹ lưỡng và thử nghiệm mã trước khi triển khai. Họ cũng khuyến nghị thực hiện các biện pháp bảo mật bổ sung, chẳng hạn như phân tích mã và thử nghiệm thâm nhập, để xác định và giảm thiểu các lỗ hổng tiềm ẩn.

3 Tối ưu hóa các quy trình làm việc trên GitHub với AI tạo sinh sử dụng Amazon Bedrock và MCP | Amazon Web Services

🇬🇧 *Streamline GitHub workflows with generative AI using Amazon Bedrock and MCP | Amazon Web Services*

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

> Amazon Web Services (AWS) đã giới thiệu một tích hợp mới cho phép người dùng tối ưu hóa các quy trình làm việc trên GitHub với sự giúp đỡ của AI tạo sinh. Tích hợp này kết hợp Amazon Bedrock, một dịch vụ được quản lý cho AI tạo sinh, với MCP (Model Canvas Platform), một nền tảng để xây dựng và triển khai các model học máy. Với tích hợp này, các developer có thể tận dụng khả năng của AI tạo sinh để tự động hóa và tối ưu hóa các quy trình làm việc trên GitHub. Điều này có thể bao gồm các nhiệm vụ như tạo mã, sửa lỗi và phát triển tính năng. Bằng cách sử dụng Amazon Bedrock và MCP, các developer có thể tạo và triển khai các model AI tùy chỉnh có thể tương tác với các kho lưu trữ GitHub của họ, giúp việc quản lý và duy trì mã nguồn trở nên dễ

dàng hơn. Tích hợp này nhằm cải thiện hiệu suất và năng suất của các developer bằng cách cung cấp một trải nghiệm tự động hóa và liền mạch cho các quy trình làm việc trên GitHub. Bằng cách tận dụng sức mạnh của AI tạo sinh, các developer có thể tập trung vào các nhiệm vụ cấp cao hơn và giảm thời gian dành cho các nhiệm vụ nhàm chán và lặp đi lặp lại.

4

AI đang được sử dụng để hồi sinh giọng nói của các phi công đã qua đời

🇬🇧 *AI is being used to resurrect the voices of dead pilots*

📰 TechCrunch AI 🔗 Đọc bài viết →

Hội đồng An toàn Giao thông Quốc gia (NTSB) đã tạm thời hạn chế truy cập vào hệ thống hồ sơ của mình sau khi phát hiện ra rằng các giọng nói được tạo bởi AI của các phi công đã qua đời từ vụ tai nạn máy bay của UPS đang được lưu hành trực tuyến. Các giọng nói này được tái tạo bằng cách sử dụng tệp spectrogram từ hồ sơ tai nạn, chứa đại diện toán học của tín hiệu âm thanh từ máy ghi âm giọng nói buồng lái. Một YouTuber, Scott Manley, đã gợi ý rằng có thể tái tạo âm thanh từ dữ liệu spectrogram. Sử dụng các công cụ AI như Codex, các cá nhân đã có thể tạo ra các bản gần□ của âm thanh máy ghi âm giọng nói buồng lái từ bản ghi và tệp spectrogram công khai. NTSB đã khôi phục lại quyền truy cập công khai vào hệ thống hồ sơ nhưng đã đóng 42 cuộc điều tra, bao gồm cả cuộc điều tra liên quan đến Chuyến bay 2976 của UPS, để xem xét lại.

5

Tiến tới Tốc độ ánh sáng trong Tạo văn bản với Mô hình ngôn ngữ khuếch tán Nemotron-Labs

🇬🇧 *Towards Speed-of-Light Text Generation with Nemotron-Labs Diffusion Language Models*

📰 Hugging Face Blog 🔗 Đọc bài viết →

Các nhà nghiên cứu tại Nemotron-Labs đã phát triển một loại mô hình ngôn ngữ mới gọi là Nemotron-Labs Diffusion, nhằm cải thiện tốc độ và hiệu quả của việc tạo văn bản. Các mô hình ngôn ngữ lớn truyền thống (LLMs) tạo văn bản một token tại một thời điểm, điều này có thể chậm và đòi hỏi nhiều bộ nhớ, đặc biệt là đối với các ứng dụng nhạy cảm về độ trễ. Nemotron-Labs Diffusion giới thiệu một phương pháp mới gọi là mô hình ngôn ngữ khuếch tán (DLMs), tạo ra nhiều token song song và sau đó tinh chỉnh chúng trong nhiều bước. Phương pháp

này có thể tận dụng sức mạnh tính toán của GPU hiện đại và mang lại lợi ích hiệu suất đáng kể. Mô hình này cũng cho phép sửa đổi các token được tạo, khiến nó phù hợp hơn với các nhiệm vụ như sửa đổi văn bản hiện có và giải quyết các mục tiêu điền vào giữa. Nemotron-Labs Diffusion bao gồm các mô hình văn bản ở các quy mô khác nhau, bao gồm 3B, 8B và 14B, cũng như một mô hình ngôn ngữ-vision, tất cả đều có sẵn dưới các giấy phép thân thiện với thương mại. Mô hình này hỗ trợ ba chế độ tạo: tự hồi quy, khuếch tán và tự suy đoán, có thể dễ dàng chuyển đổi giữa chúng tại thời điểm triển khai. So với các mô hình hiện có, Nemotron-Labs Diffusion 8B đạt được độ chính xác trung bình cải thiện 1,2%.

6

Các giao dịch công nghệ trong Ngày tưởng niệm: Sony, Apple, Beats (2026)

🇬🇧 *Memorial Day Tech Deals: Sony, Apple, Beats (2026)*

📰 Wired 🔗 Đọc bài viết →

Những người đam mê công nghệ, hãy vui mừng. Các chương trình giảm giá vào ngày lễ tưởng niệm đã đến, và cùng với đó, các sản phẩm công nghệ hàng đầu đang được giảm giá. Tai nghe không dây WH-1000XM5 của Sony, một phiên bản tiền nhiệm của model yêu thích của chúng tôi, hiện có sẵn với giá thấp nhất trong một thời gian, chỉ thiếu 5 đô la so với mức giá thấp nhất từ trước đến nay. Các ưu đãi khác bao gồm loa Bluetooth tốt nhất của Apple, một sản phẩm thời trang, di động và có âm thanh ấn tượng, hiện có sẵn với hầu hết màu sắc tại mức giá giảm. Laptop tốt nhất, nổi tiếng với sức mạnh, thời lượng pin và thiết kế tinh tế, cũng đang được giảm giá, mặc dù nó đã ở mức giá này trong vài tuần. Ngoài ra, các sản phẩm được WIRED khuyên dùng như earbuds, pin dự phòng, máy đi bộ và bộ khởi động di động cũng đang được giảm giá. Các ưu đãi khác bao gồm bộ chuyển đổi du lịch, tai nghe chơi game, mũ trị liệu tóc bằng ánh sáng đỏ và earbuds mở. Những chương trình giảm giá này là cơ hội tuyệt vời để nâng cấp thiết bị công nghệ của bạn mà không cần phải chi tiêu quá nhiều.

7

Tổng chưởng lý Texas kiện Meta vì cáo buộc WhatsApp không cung cấp mã hóa từ đầu đến đầu

🇬🇧 *Texas AG sues Meta over claims that WhatsApp doesn't provide end-to-end encryption*

📰 Ars Technica 🔗 Đọc bài viết →

Tổng chưởng lý Texas đã đệ đơn kiện Meta, công ty mẹ của WhatsApp, với cáo buộc rằng dịch vụ nhắn tin này không cung cấp mã hóa end-to-end (E2EE) như đã tuyên bố. WhatsApp đã khẳng định rằng họ cung cấp E2EE mạnh mẽ, nghĩa là các tin nhắn được mã hóa trên thiết bị của người gửi và chỉ có thể được đọc bởi người nhận dự kiến. Tuy nhiên, Tổng chưởng lý Texas cho rằng Meta có thể và đã đọc nội dung không được mã hóa của các tin nhắn WhatsApp. Vụ kiện này trích dẫn một bài báo của Bloomberg làm bằng chứng, đã đưa tin rằng Bộ Thương mại Hoa Kỳ, Văn phòng An ninh Công nghiệp đã đóng một cuộc điều tra về cáo buộc rằng Meta có thể truy cập vào các tin nhắn WhatsApp được mã hóa. Meta đã phủ nhận các cáo buộc, gọi chúng là "không có căn cứ" và cam kết sẽ chống lại vụ kiện tại tòa. Các nhà phê bình, bao gồm các nhà công nghệ và chuyên gia mã hóa, đã lưu ý rằng thiếu hỗ trợ thực cho các cáo buộc, với một số gợi ý rằng một kỹ thuật đảo ngược kỹ lưỡng của WhatsApp sẽ là cần thiết để xác nhận hoặc phủ nhận các cáo buộc.

8

Từ Dòng token đến Dòng agent

🇬🇧 *From Token Streams to Agent Streams*

📰 LangChain Blog 🔗 Đọc bài viết →

Một nền tảng phát trực tuyến mới đã được phát triển cho các tác nhân AI phức tạp có thể thực hiện nhiều nhiệm vụ, chẳng hạn như lập kế hoạch, ủy quyền và sản xuất văn bản hoặc phương tiện. Các API phát trực tuyến trước đây được thiết kế cho các cuộc gọi model đơn lẻ và không thể xử lý sự phức tạp của các tác nhân này. Nền tảng mới sử dụng các sự kiện ứng dụng thay vì các khối thô, cho phép kiểm soát nhiều hơn về những gì được phát trực tuyến và cách nó được trình bày. Mỗi sự kiện được gán loại và gắn thẻ với nguồn gốc của nó trong cây tác nhân, cho phép các ứng dụng đăng ký vào các phần cụ thể của quy trình làm việc của tác nhân. Điều này cho phép phát trực tuyến hiệu quả hơn và giảm tải công việc cho các nhà phát triển. Lớp phát trực tuyến mới được thiết kế để xử lý các tác nhân có hình dạng đồ thị, sử dụng công cụ, có trạng thái, có thể bị gián đoạn và đa phương thức chạy trên các backend và frontend. Nó sử dụng một phong bì sự kiện chung và tách các kênh khỏi không gian tên, làm cho nó dễ dàng xác định phần của tác nhân tạo ra sự kiện. Lớp phát trực tuyến hiển thị các dự án có kiểu trên dòng sự kiện, cho phép các ứng dụng yêu cầu nội dung cụ thể, chẳng hạn như văn bản, lập luận hoặc các đối số gọi công

> cụ, thay vì lặp lại các sự kiện giao thức thô. Điều này cho phép việc kết xuất đầu ra model phức tạp hiệu quả và chính xác hơn.

⚡ TIPS & TRICKS CHO DEV

⚡ Sử dụng LangGraph

Vấn đề: Khó khăn trong việc phối hợp nhiều AI agents để xử lý task phức tạp.
Cách làm: Sử dụng LangGraph để tạo ra một mạng lưới các AI agents, ví dụ như sử dụng lệnh `langgraph init` để khởi tạo dự án.
Đánh giá: Hiệu quả trong việc tăng tốc độ xử lý và giảm thiểu sai sót, nên dùng khi cần phối hợp nhiều AI agents.

⚡ Tích hợp CrewAI

Vấn đề: Thiếu sự linh hoạt trong việc điều chỉnh và cập nhật các AI agents.
Cách làm: Tích hợp CrewAI vào dự án để có thể điều chỉnh và cập nhật các AI agents một cách linh hoạt, ví dụ như sử dụng lệnh `crewai update`.
Đánh giá: Giúp tăng sự linh hoạt và khả năng thích ứng của hệ thống, nên dùng khi cần thường xuyên cập nhật và điều chỉnh AI agents.

⚡ Áp dụng AutoGen

Vấn đề: Tốn thời gian và công sức để tạo ra các AI agents mới.
Cách làm: Sử dụng AutoGen để tự động tạo ra các AI agents mới, ví dụ như sử dụng lệnh `autogen create`.
Đánh giá: Giúp giảm thiểu thời gian và công sức, nên dùng khi cần tạo ra nhiều AI agents mới một cách nhanh chóng.

📚 BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để phát triển ứng dụng hiệu quả, các nhà phát triển cần tối ưu hóa 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, đảm bảo ứng dụng hoạt động mượt mà.
3. Ví dụ, sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và tăng tốc độ xử lý. Ví dụ code: `model = transformers.AutoModelForSequenceClassification.from_pretrained('model_name', num_labels=8)`.
4. 💡 Tip: Sử dụng thư viện như Hugging Face Transformers để tối ưu hóa mô hình và giảm thiểu chi phí tính toán.

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

Nguồn: Google News · Groq AI