BIS - MT · 17/05/2026

🤖 Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

✨ *"The strength of the team is each individual member. The strength of each member is the team."*

↪ Sức mạnh của đội nhóm là từng thành viên. Sức mạnh của từng thành viên chính là đội nhóm.

— Phil Jackson

💡 *Cá nhân và tập thể bổ trợ nhau — đội nhóm mạnh giúp từng người phát triển tốt hơn, và mỗi người mạnh lại tăng cường sức mạnh chung.*

📰 TIN TỨC NỔI BẬT

1

Claude Code đấu với GitHub Copilot 2026: SWE-bench, Giá cả [Đã thử nghiệm]

🇬🇧 *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 đưa vào thử nghiệm về hiệu suất và giá cả. SWE-bench, một công cụ benchmark cho phát triển phần mềm, đã được sử dụng để đánh giá hai trợ lý coding được hỗ trợ bởi AI này. Kết quả cho thấy Claude Code vượt trội hơn GitHub Copilot ở một số lĩnh vực, bao gồm code completion, code review và debugging. Claude Code đã thể hiện tính năng code completion chính xác và hiệu quả hơn, với tỷ lệ thành công cao hơn trong việc hoàn thành các đoạn code snippet. Ngoài ra, tính năng code review của nó được đánh giá là hiệu quả hơn trong việc xác định và đề xuất cải tiến cho code. Tuy nhiên, giá cả của hai dịch vụ khác nhau đáng kể, với GitHub Copilot cung cấp một lựa chọn phải chăng hơn cho cá nhân và các team nhỏ. Nghiên cứu cho thấy Claude Code có thể là lựa chọn tốt hơn cho các developer yêu cầu hỗ trợ coding hiệu suất cao, đặc biệt trong các dự án quy mô lớn. Mặt khác, mức giá thấp hơn của GitHub Copilot có thể khiến nó trở thành một lựa chọn hấp dẫn hơn cho các team nhỏ hơn hoặc cá nhân có ngân sách hạn chế.

2 Ai cũng muốn tham gia vibe coding — và Google cũng không ngoại lệ với Stitch, dự án kế nhiệm Jules

🇬🇧 *Everyone's looking to get in on vibe coding — and Google is no different with Stitch, its follow-up to Jules*

📰 VentureBeat 🔗 Đọc bài viết →

> Google đang gia nhập không gian vibe coding với Stitch, dự án phát triển mới nhất của mình. Vibe coding, một khái niệm tương đối mới, tập trung vào việc tạo ra trải nghiệm coding nhập vai và tương tác hơn. Stitch là dự án kế nhiệm của Jules, dự án trước đây của Google, cho thấy sự đầu tư liên tục của công ty vào lĩnh vực này. Thông tin chi tiết về Stitch hiện còn hạn chế, nhưng việc ra mắt nó báo hiệu sự quan tâm của Google trong việc mở rộng các công cụ và nền tảng coding của mình. Gã khổng lồ công nghệ này có thể đang tìm cách nâng cao cách các developer học và tương tác với code, có thể kết hợp các yếu tố phát triển game và kể chuyện tương tác vào quá trình coding. Khi xu hướng vibe coding ngày càng phát triển, việc Google tham gia vào không gian này là đáng chú ý. Với Stitch, công ty có thể đang định vị mình để hỗ trợ tốt hơn nhu cầu của các developer hiện đại và cung cấp trải nghiệm coding hấp dẫn hơn. Thông tin thêm về các tính năng và khả năng của Stitch dự kiến sẽ được công bố trong những ngày tới.

3 LangChain ra mắt Deep Agents: Một runtime có cấu trúc để lập kế hoạch, quản lý bộ nhớ và cô lập ngữ cảnh trong các AI agent đa bước

🇬🇧 *LangChain Releases Deep Agents: A Structured Runtime for Planning, Memory, and Context Isolation in Multi-Step AI Agents*

📰 MarkTechPost 🔗 Đọc bài viết →

> LangChain đã công bố phát hành Deep Agents, một runtime có cấu trúc để lập kế hoạch, quản lý bộ nhớ và cô lập ngữ cảnh trong các AI agent đa bước. Công nghệ này nhằm mục đích cải thiện hiệu suất và độ tin cậy của các hệ thống AI phức tạp bằng cách cung cấp một framework để quản lý nhiều tác vụ và duy trì ngữ cảnh qua các bước khác nhau. Deep Agents được thiết kế để xử lý các tác vụ yêu cầu lập kế hoạch, quản lý bộ nhớ và cô lập ngữ cảnh, chẳng hạn như hệ thống hỏi đáp, chatbot và các ứng dụng ra quyết định. Runtime có cấu trúc cho phép các developer xây dựng và triển khai các AI agent đa bước dễ dàng và hiệu quả hơn. Các tính năng chính của Deep Agents bao gồm hỗ trợ lập kế hoạch, quản lý bộ nhớ và cô lập ngữ cảnh, cũng như tích hợp với các AI framework và thư viện phổ biến. Công nghệ này dự

kiến sẽ mang lại lợi ích cho các developer làm việc trên các dự án AI phức tạp, cho phép họ tạo ra các hệ thống tinh vi và đáng tin cậy hơn. Với việc phát hành Deep Agents, LangChain tiếp tục thúc đẩy lĩnh vực phát triển và triển khai AI.

4 Những góc nhìn mới ấn tượng về vụ thử bom nguyên tử đầu tiên

🇬🇧 *Striking New Views of the First Atomic Bomb Test*

📰 IEEE Spectrum 🔗 Đọc bài viết →

Những bức ảnh mới được phát hiện về vụ thử bom nguyên tử Trinity cung cấp một minh họa trực quan ấn tượng về sức mạnh khủng khiếp của sự kiện này. Các hình ảnh, được Berlyn Brixner chụp bằng hai máy quay phim Mitchell, cho thấy quả cầu lửa giãn nở nhanh chóng, với vụ nổ đã rộng hàng trăm mét chỉ 0,016 giây sau khi kích nổ. Các bức ảnh, được chụp từ một boong-ke, cũng tiết lộ ánh sáng và nhiệt độ dữ dội phát ra từ vụ nổ, mạnh hơn dự đoán vài lần. Mặc dù cường độ áp đảo của vụ nổ, các máy quay đã ghi lại được một bức tranh hoàn chỉnh đáng kinh ngạc về sự kiện, với hơn 100.000 khung hình có sẵn để phân tích. Các bức ảnh và đoạn phim đã vô cùng quý giá đối với các nhà khoa học, cho phép họ đo lường và mô tả hành vi của quả cầu lửa và các hiệu ứng nhìn thấy khác với chi tiết chính xác. Nỗ lực chụp ảnh là một thành công lớn, mặc dù chỉ có 11 trong số 52 máy quay tạo ra hình ảnh đạt yêu cầu, và đã cung cấp một cái nhìn độc đáo về một trong những sự kiện quan trọng nhất trong lịch sử.

5 Cisco công bố doanh thu kỷ lục và sa thải 4.000 nhân viên trong cùng một ngày

🇬🇧 *Cisco announces record revenue and 4,000 layoffs in the same day*

📰 Ars Technica 🔗 Đọc bài viết →

Cisco, một công ty công nghệ hàng đầu, đã công bố doanh thu kỷ lục 15,8 tỷ USD trong quý tài chính Q3 năm 2026, tăng 12% so với cùng kỳ năm trước. Tuy nhiên, bất chấp thành công này, công ty đang trải qua một nỗ lực tái cấu trúc đáng kể, bao gồm việc sa thải khoảng 4.000 nhân viên, chiếm chưa đến 5% tổng lực lượng lao động. Việc sa thải được cho là do sự phát triển của Trí tuệ nhân tạo (AI) và nhu cầu của công ty phải tập trung vào các lĩnh vực có nhu cầu và tạo ra giá trị dài hạn mạnh nhất. Cisco có kế hoạch đầu tư vào các lĩnh vực như

silicon, quang học, bảo mật và AI, đồng thời sẽ hỗ trợ các nhân viên bị ảnh hưởng, bao gồm thanh toán tiền thưởng theo tỷ lệ và tiếp cận các dịch vụ tìm việc làm cũng như cơ hội học tập cá nhân hóa. Công ty dự kiến sẽ ghi nhận khoản chi phí trước thuế lên tới 1 tỷ USD liên quan đến việc sa thải, với 450 triệu USD sẽ được ghi nhận trong Q4 FY '26 và phần còn lại trong FY '27.

6

Xây dựng một accessibility agent đa năng — và những gì chúng tôi học được trong quá trình này

🇬🇧 *Building a general-purpose accessibility agent—and what we learned in the process*

📰 GitHub Blog 🔗 Đọc bài viết →

GitHub đang thử nghiệm một accessibility agent đa năng để cải thiện khả năng tiếp cận của nền tảng. Agent này nhằm mục đích tự động đánh giá các thay đổi đối với code front-end và xác định các vấn đề tiềm ẩn có thể gây cản trở cho người dùng sử dụng công nghệ hỗ trợ. Agent đã xem xét 3.535 pull request, giải quyết 68% các vấn đề được tìm thấy. Năm loại vấn đề hàng đầu mà agent đã xác định bao gồm các vấn đề về layout và styling, các vấn đề liên quan đến accessibility và các vấn đề với điều hướng bằng bàn phím. Những vấn đề này có thể tạo ra ma sát và rào cản cho người dùng dựa vào công nghệ hỗ trợ. Agent là một ví dụ về cách GitHub đang sử dụng trí tuệ nhân tạo và machine learning để cải thiện trải nghiệm developer và làm cho nền tảng dễ tiếp cận hơn. Công ty có kế hoạch chia sẻ những thành công và bài học kinh nghiệm từ thử nghiệm này, với mục tiêu giúp những người khác cải thiện accessibility trong các quy trình phát triển phần mềm của riêng họ.

7

The Download: "Nhà máy" phim ngắn AI của Trung Quốc và các mục tiêu y tế bị bỏ lỡ của WHO

🇬🇧 *The Download: China's AI drama factory and the WHO's missing health targets*

📰 MIT Tech Review 🔗 Đọc bài viết →

Dưới đây là tóm tắt tin tức công nghệ khoảng 150-200 từ: Ngành công nghiệp phim ngắn của Trung Quốc đang trải qua một sự chuyển đổi đáng kể, được thúc đẩy bởi việc sử dụng trí tuệ nhân tạo (AI) ngày càng tăng. Các phim ngắn do AI tạo ra đang được sản xuất với tốc độ chưa từng có, với trung bình 470 phim được phát hành mỗi ngày trong tháng 1. Sự thay đổi này đã rút ngắn thời gian sản xuất từ vài tháng

xuống còn vài tuần và giảm chi phí tới 90%. Định dạng này đang mở rộng ra toàn cầu, định hình lại công việc của các nhà biên

8 Cách chúng tôi sử dụng Sourcegraph và một Slack bot để phát hiện lỗ hổng và phản ứng nhanh chóng

🇬🇧 *How we're using Sourcegraph and a Slack bot to detect vulnerabilities and react quickly*

📰 Sourcegraph Blog 🔗 Đọc bài viết →

> Một công ty công nghệ đang sử dụng Sourcegraph và một bot Slack để phát hiện và phản hồi hiệu quả các điểm yếu tiềm năng. Bot Slack tự động phân loại từng lời khuyên trên GitHub, đăng một thông báo trên kênh để thu hút sự chú ý đến vấn đề. Khi một thành viên trong nhóm phản hồi thông báo, bot kích hoạt một đường ống công việc toàn diện. Điều này bao gồm chạy các truy vấn phát hiện, tạo nội dung blog, tạo bản nháp truyền thông xã hội và sản xuất một video demo 35 giây. Khi đường ống hoàn thành, thành viên trong nhóm chỉ còn nhiệm vụ xem xét nội dung được tạo để đảm bảo độ chính xác. Quá trình được tối ưu hóa này cho phép công ty nhanh chóng xác định và giải quyết các điểm yếu tiềm năng, cho phép hành động nhanh chóng được thực hiện để giảm thiểu mọi rủi ro.

⚡ TIPS & TRICKS CHO DEV

⚡ Cài đặt Ollama

Vấn đề: Máy tính không có kết nối internet để sử dụng LLM.

Cách làm: Sử dụng Ollama, cài đặt bằng lệnh `pip install ollama`. Ví dụ, chạy lệnh `ollama --model small` để khởi động mô hình nhỏ.

Đánh giá: Hiệu quả cao, phù hợp khi cần sử dụng LLM offline.

⚡ Tối ưu LM Studio

Vấn đề: LM Studio tiêu tốn nhiều tài nguyên hệ thống.

Cách làm: Tối ưu hóa LM Studio bằng cách điều chỉnh kích thước mô hình và số lượng nhân viên xử lý. Ví dụ, chạy lệnh `lm-studio --model-size small --num-workers 2`.

Đánh giá: Giúp giảm tải hệ thống, tăng hiệu suất.

⚡ Sử dụng Prompt

Vấn đề: LLM không hiểu rõ yêu cầu người dùng.

Cách làm: Sử dụng câu prompt rõ ràng, như "Tóm tắt nội dung của đoạn văn này". Ví dụ, nhập prompt vào Ollama: `ollama --prompt "Tóm tắt đoạn văn"` để nhận

được kết quả chính xác.
Đánh giá: Cải thiện độ chính xác của LLM, giúp người dùng nhận được thông tin hữu ích.

📚 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 đảm bảo ứng dụng AI của họ hoạt động hiệu quả và tiết kiệm chi phí. Điều này đặc biệt quan trọng khi triển khai mô hình AI trên quy mô lớn.

2. Việc tối ưu hóa hiệu năng LLM giúp giảm thiểu thời gian phản hồi và tăng cường trải nghiệm người dùng, trong khi tối ưu chi phí giúp giảm thiểu chi phí vận hành và bảo trì.

3. Ví dụ, sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm thiểu kích thước mô hình và tăng cường hiệu năng.

4. 💡 Tip hoặc bước tiếp theo: Để tối ưu hóa hiệu năng LLM, hãy thử sử dụng các kỹ thuật như quantization, pruning, và knowledge distillation để giảm thiểu kích thước mô hình và tăng cường hiệu năng.