

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Biết người biết ta, trăm trận trăm thắng.”

— Tôn Tử

Hiểu rõ bản thân và đối tác/đối thủ là nền tảng của mọi quyết định sáng suốt trong công việc lẫn cuộc sống.

TIN TỨC NỔI BẬT

1

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

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 thử nghiệm về hiệu suất và giá cả. Công cụ SWE-bench, một công cụ chuẩn để đánh giá kỹ thuật phần mềm, đã được sử dụng để đánh giá hai trợ lý mã hóa được hỗ trợ bởi AI. Kết quả cho thấy Claude Code vượt trội so với GitHub Copilot ở nhiều lĩnh vực, bao gồm hoàn thành mã và gỡ lỗi. Giá cả của hai dịch vụ cũng được so sánh. GitHub Copilot cung cấp một phiên bản thử nghiệm miễn phí, nhưng mô hình đăng ký của nó bắt đầu từ 10 đô la mỗi tháng. Mặt khác, Claude Code cung cấp một kế hoạch miễn phí với các tính năng hạn chế, cũng như một kế hoạch trả phí bắt đầu từ 15 đô la mỗi tháng. Kế hoạch trả phí của Claude Code bao gồm các tính năng bổ sung như mô hình tùy chỉnh và hỗ trợ ưu tiên. Tổng thể, so sánh này gợi ý rằng Claude Code có thể là một lựa chọn tiết kiệm hơn cho các nhà phát triển yêu cầu hỗ trợ mã hóa tiên tiến. Tuy nhiên, kết quả của thử nghiệm SWE-bench nên được xem như một chỉ số chung về hiệu suất, vì nhu cầu cụ thể của từng nhà phát triển có thể khác nhau.

2

Multi-Agent Warehouse AI Command Layer: Nâng tầm Operational Excellence và Supply Chain Intelligence

Multi-Agent Warehouse AI Command Layer Enables Operational Excellence and Supply Chain Intelligence

NVIDIA Developer [Đọc bài viết →](#)

NVIDIA đã phát triển một Lớp lệnh AI Nhà kho Đa tác nhân, một công nghệ tiên tiến được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng trong các nhà kho. Giải pháp đổi mới này sử dụng trí tuệ nhân tạo (AI) để tối ưu hóa hoạt động nhà kho, cho phép các doanh nghiệp tối ưu hóa hậu cần và đưa ra quyết định dựa trên dữ liệu. Lớp lệnh AI Nhà kho Đa tác nhân là một trung tâm điều hành tích hợp nhiều tác nhân AI, mỗi tác nhân chịu trách nhiệm cho một nhiệm vụ cụ thể, chẳng hạn như quản lý hàng tồn kho, thực hiện đơn hàng và phân bổ tài nguyên. Các tác nhân này làm việc cùng nhau để phân tích dữ liệu thời gian thực từ các nguồn khác nhau, bao gồm cảm biến, camera và thiết bị IoT, để cung cấp một cái nhìn tổng thể về hoạt động nhà kho. Bằng cách tận dụng AI và học máy, Lớp lệnh AI Nhà kho Đa tác nhân có thể xác định các điểm nghẽn, dự đoán nhu cầu và tối ưu hóa quy trình làm việc, dẫn đến năng suất lao động được cải thiện, chi phí giảm và sự hài lòng của khách hàng được nâng cao. Công nghệ này có tiềm năng cách mạng hóa cách thức hoạt động của các nhà kho, cho phép các doanh nghiệp đạt được sự xuất sắc trong hoạt động và thu được những thông tin quý giá về chuỗi cung ứng của họ.

3

Claude vs. ChatGPT 2026: Thống kê đối đầu, những con số đằng sau cuộc 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 Claude và ChatGPT đã làm sáng tỏ sự khác biệt về hiệu suất giữa hai model AI này. Theo dữ liệu, Claude đã vượt trội so với ChatGPT trong một số lĩnh vực quan trọng. Về luồng hội thoại, Claude đã đạt được tỷ lệ thành công 92%, trong khi ChatGPT đạt được 85%. Khi nói đến việc trả lời các câu hỏi tiếp theo, Claude đã trả lời đúng 88% thời gian, so với 78% của ChatGPT. Về cơ sở kiến thức, cơ sở dữ liệu của Claude chứa 1,5 tỷ tham số, trong khi ChatGPT có 1 tỷ. Tuy nhiên, dữ liệu đào tạo của ChatGPT đa dạng hơn, với một loạt các chủ đề và nguồn rộng lớn hơn. Dữ liệu đào tạo của Claude, mặt khác, tập trung hơn vào hội thoại. Thống kê cũng cho thấy Claude có khả năng xử lý các truy vấn phức tạp tốt hơn, với tỷ lệ thành công 95% trong lĩnh vực này. Ngược lại, tỷ lệ thành công của ChatGPT đối với các truy vấn phức tạp là 85%. Tổng thể, dữ liệu cho thấy Claude có một lợi thế nhỏ so với ChatGPT về hiệu suất hội thoại và xử lý các truy vấn phức tạp.

4

AI hedge fund Situational Awareness bán public portfolio, vẫn giữ cổ phần Anthropic

AI hedge fund Situational Awareness may have sold its public portfolio, but it still has its Anthropic shares

TechCrunch AI

[Đọc bài viết →](#)

Situational Awareness, một quỹ đầu tư chứng khoán do 25 tuổi Leopold Aschenbrenner thành lập, đã bán phần lớn danh mục đầu tư chứng khoán công khai của mình cho Citadel, một quỹ đầu tư chứng khoán do Ken Griffin lãnh đạo, sau khi lỗ đáng kể trong tháng trước. Quỹ này, đã mang lại lợi nhuận 439% cho năm qua đến tháng 6, đã có tài sản quản lý lên tới 45 tỷ đô la trước khi các vị trí của nó bắt đầu giảm trong bối cảnh suy giảm rộng rãi hơn trong các khoản đầu tư cơ sở hạ tầng AI. Aschenbrenner, một nhà nghiên cứu OpenAI cũ, đã đầu tư mạnh vào các công ty như SK Hynix, Bloom Energy và Nebius Group, đã giảm hơn 30% trong tháng trước. Những tổn thất của quỹ đã được làm trầm trọng thêm bởi đòn bẩy, một chiến lược quỹ đầu tư chứng khoán phổ biến. Mặc dù lỗ, Aschenbrenner vẫn giữ thái độ lạc quan, mời khách hàng cam kết vốn mới trong một bức thư ngày 24 tháng 7. Tuy nhiên, lời mời này không thành công, và tài sản của quỹ đã giảm xuống còn khoảng 10 tỷ đô la. Đáng chú ý, Situational Awareness vẫn nắm giữ một cổ phần đáng kể trong công ty tư nhân Anthropic, được định giá 5 tỷ đô la, dự kiến sẽ được niêm yết công khai vào tháng 10 và có thể đạt được định giá thậm chí cao hơn.

5

Đẩy xa price-performance frontier với GPT-5.6

Advancing the price-performance frontier with GPT-5.6

Simon Willison

[Đọc bài viết →](#)

OpenAI đã công bố giảm giá đáng kể cho các model GPT-5.6 của mình, với GPT-5.6 Terra giảm 20% và GPT-5.6 Luna giảm 80%. Việc giảm giá này được cho là nhờ vào nỗ lực tối ưu hóa của GPT-5.6 Sol, một thành phần sử dụng trí tuệ nhân tạo (AI) để tối ưu hóa hiệu suất và hiệu quả của model. GPT-5.6 Sol đã đạt được điều này bằng cách tối ưu hóa cân bằng tải, suy luận và tính toán model, dẫn đến giảm 20% chi phí phục vụ từ đầu đến cuối. Đặc biệt, model Luna đã trở thành một lựa chọn tiết kiệm hơn, với giá thấp hơn so với Gemini 3.1 Flash-Lite của Google và Claude Haiku 4.5 của Anthropic. Kết quả là, các nhà phát triển (developer) có thể xem xét chuyển sang model Luna tiết kiệm hơn,

nếu được chứng minh bởi việc chuyển đổi của tác giả từ Gemini 3.1 Flash-Lite sang Luna.

6

Vượt xa passive RAG: Cách triển khai active memory reconstruction cho 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 gọi là MRAgent để cải thiện khả năng lập luận của các tác nhân AI trong lịch sử hội thoại dài hạn. Cách tiếp cận "truy xuất rồi lập luận" hiện tại, nơi các tác nhân AI truy xuất thông tin từ bộ nhớ và sau đó lập luận trên đó, bị hạn chế và tạo ra tiếng ồn trong ngữ cảnh. MRAgent thay thế cách tiếp cận tĩnh này bằng một cơ chế tái tạo bộ nhớ chủ động, tích hợp tái tạo bộ nhớ vào quá trình lập luận của Large Language Model (LLM). Điều này cho phép tác nhân dần dần phát triển bộ nhớ của mình dựa trên bằng chứng và kinh nghiệm tích lũy. Các thí nghiệm đã chỉ ra rằng MRAgent có độ chính xác cao và hiệu quả về tài nguyên, đạt được sự cải thiện hiệu suất 23% trên các nhiệm vụ lập luận phức tạp. Nó cũng giảm tiêu thụ token prompt và chi phí thời gian chạy, làm cho nó phù hợp với các ứng dụng thực tế với tương tác dài hạn. Framework MRAgent giải quyết các hạn chế của các đường ống truy xuất thụ động hiện tại, tách biệt lập luận khỏi truy cập bộ nhớ, tạo ra nút thắt khi mở rộng đến các trường hợp sử dụng doanh nghiệp.

7

AGI không phải Multimodal

AGI Is Not Multimodal

The Gradient

[Đọc bài viết →](#)

Những tiến bộ gần đây trong các mô hình AI tạo sinh đã khiến một số người tin rằng Trí tuệ Nhân tạo Tổng quát (AGI) đang sắp xảy ra. Tuy nhiên, giả định này dựa trên khả năng của các mô hình có thể mở rộng hiệu quả trên phần cứng hiện có, chứ không phải là một cách tiếp cận có suy nghĩ về trí tuệ. Phương pháp đa mô thức, kết hợp nhiều mô thức để tạo ra một AI tổng quát, được coi là một con đường tiềm năng đến AGI. Tuy nhiên, chiến lược này không thể thành công trong thời gian gần, vì nó không giải quyết các khía cạnh cơ bản của trí tuệ con người, chẳng hạn như lý luận cảm biến và phối hợp xã hội. Một AGI

thực sự phải là tổng quát trên tất cả các lĩnh vực, bao gồm cả thực tại vật lý. Điều này đòi hỏi một hình thức trí tuệ cơ bản được đặt trong một mô hình thế giới vật lý. Các mô hình Ngôn ngữ Lớn (LLM) hiện tại không đạt được điều này, vì chúng đang học "túi các quy tắc gần đúng" để dự đoán token, chứ không phải là một sự hiểu biết sâu sắc về thực tại. Mặc dù LLM có thể xuất hiện như có sự hiểu biết giống con người về thế giới, nhưng điều này là bề mặt và không phản ánh một sự hiểu biết thực sự về thế giới vật lý.

8

Kế hoạch biến Montana thành experimental medical hub đã được đẩy mạnh

Montana's plan to become an experimental medical hub just pushed forward

MIT Tech Review

[Đọc bài viết →](#)

Montana đã thực hiện một bước tiến đáng kể để trở thành một trung tâm y tế thí nghiệm, với luật "quyền được thử" của tiểu bang này được mở rộng để cho phép bất kỳ công ty công nghệ sinh học nào có thuốc thử nghiệm bán nó cho người tiêu dùng. Các công ty có thể nộp đơn xin phê duyệt từ một hội đồng xem xét với giá 12.500 đô la, sau đó họ có thể tự đặt giá và bán thuốc thông qua các phòng khám điều trị thử nghiệm. Luật này độc đáo ở chỗ nó không giới hạn việc tiếp cận thuốc cho những người bị bệnh nan y, mà cho phép bất kỳ ai cung cấp sự đồng ý thông tin và có thể trả tiền để tiếp cận các phương pháp điều trị thử nghiệm. Bộ Y tế và Dịch vụ Nhân sinh của tiểu bang đã hoàn thiện các quy tắc để thực hiện luật này, bao gồm việc xem xét nghiêm ngặt bởi một hội đồng các chuyên gia y tế và một nhà đạo đức. Những người ủng hộ luật này, bao gồm cả những người đam mê tuổi thọ, nhấn mạnh rằng quá trình này sẽ được thực hiện một cách có trách nhiệm. Tuy nhiên, các chuyên gia đã bày tỏ lo ngại về tiềm năng gây hại khi bán các phương pháp điều trị chưa được chứng minh cho những người không có sự giám sát từ Cơ quan Quản lý Thực phẩm và Dược phẩm Hoa Kỳ.

TIPS & TRICKS CHO DEV

Tích hợp File System

Vấn đề: Kết nối AI với file system để truy xuất dữ liệu.

Cách làm: Sử dụng MCP để kết nối Claude Desktop với file system, cho phép truy xuất và chỉnh sửa file trực tiếp. Ví dụ: `claude-desktop --connect-file-system`.

Đánh giá: Hiệu quả khi cần chỉnh sửa file trực tiếp, nhưng đòi hỏi cấu hình chính xác.

Kết nối Git

Vấn đề: Đồng bộ hóa dữ liệu giữa AI và Git.

Cách làm: Sử dụng MCP để kết nối Claude Code với Git, cho phép đẩy và kéo code trực tiếp. Ví dụ: `claude-code --connect-git`.

Đánh giá: Hiệu quả khi cần đồng bộ hóa code, nhưng đòi hỏi cấu hình Git chính xác.

Tích hợp Database

Vấn đề: Kết nối AI với database để truy xuất dữ liệu.

Cách làm: Sử dụng MCP để kết nối Cursor với database, cho phép truy xuất và chỉnh sửa dữ liệu trực tiếp. Ví dụ: `cursor --connect-database`.

Đánh giá: Hiệu quả khi cần truy xuất dữ liệu lớn, nhưng đòi hỏi cấu hình database chính xác.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Trong việc phát triển ứng dụng AI, tối ưu chi phí và hiệu năng của các mô hình ngôn ngữ lớn (LLM) là rất quan trọng để giảm thiểu chi phí tính toán và cải thiện hiệu suất. Các dev cần biết cách tối ưu hóa để đảm bảo ứng dụng của mình hoạt động hiệu quả.

3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning để điều chỉnh mô hình cho phù hợp với nhiệm vụ cụ thể, hoặc sử dụng các công cụ như LoRA để giảm kích thước mô hình.

4. Tip hoặc bước tiếp theo: Hãy bắt đầu bằng việc phân tích yêu cầu của ứng dụng và xác định những phần của mô hình có thể được tối ưu hóa, sau đó áp dụng các kỹ thuật như pruning, quantization hoặc knowledge distillation để giảm chi phí và cải thiện hiệu năng.

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

Nguồn: Google News · Groq AI