BIS - MT · 06/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“When you reach the end of your rope, tie a knot in it and hang on.”

↪ Khi bạn đến cuối sợi dây, hãy thắt một nút và bám chặt vào.

— Franklin D. Roosevelt

Khi đã dồn đến giới hạn, hãy tìm cách giữ vững — sự kiên cường trong giây phút tuyệt vọng nhất thường là điểm khởi đầu của sự đột phá.

TIN TỨC NỔI BẬT

1

Claude 4.1 mới của Anthropic chiếm ưu thế trong các bài kiểm tra mã hóa chỉ vài ngày trước khi GPT-5 ra mắt

Anthropic’s new Claude 4.1 dominates coding tests days before GPT-5 arrives

VentureBeat Đọc bài viết →

> Anthropic đã phát hành Claude 4.1, một bản cập nhật đáng kể cho mô hình AI của họ. Trong một bài kiểm tra mã hóa gần đây, Claude 4.1 đã thể hiện hiệu suất vượt trội, vượt qua các đối thủ cạnh tranh. Thành tựu này đến chỉ vài ngày trước khi phát hành GPT-5, một mô hình AI đối thủ được phát triển bởi OpenAI, được mong đợi cao. Kết quả ấn tượng của Claude 4.1 trong bài kiểm tra mã hóa là minh chứng cho những tiến bộ mà Anthropic đã đạt được trong lĩnh vực AI. Khả năng của mô hình đã được tinh chỉnh, cho phép nó xử lý các nhiệm vụ mã hóa phức tạp một cách dễ dàng. Bản cập nhật này là một bước tiến quan trọng cho Anthropic, đặt công ty này vào vị trí là một người chơi chính trong ngành công nghiệp AI. Việc phát hành Claude 4.1 và sự ra mắt sắp tới của GPT-5 dự kiến sẽ có tác động đáng kể đến cảnh quan AI. Khi sự cạnh tranh giữa hai mô hình này trở nên gay gắt, các nhà phát triển và người dùng có thể mong đợi thấy những tiến bộ nhanh chóng trong khả năng AI. Kết quả của bài kiểm tra mã hóa là một chỉ số rõ ràng về công việc sáng tạo được thực hiện bởi Anthropic, và công ty này có khả năng sẽ tiếp tục đẩy ranh giới của những gì có thể đạt được với AI.

2 So sánh 7 mô hình ngôn ngữ lớn hàng đầu (LLMs/Hệ thống) cho mã hóa năm 2025

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

MarkTechPost Đọc bài viết →

Vào năm 2025, lĩnh vực mô hình ngôn ngữ lớn (LLMs) đã chứng kiến những tiến bộ đáng kể, đặc biệt là trong ứng dụng của chúng vào việc lập trình. Bài viết này so sánh 7 LLMs/hệ thống hàng đầu cho việc lập trình, nhấn mạnh khả năng và tính năng của chúng. Các mô hình bao gồm: 1. LLaMA (Mô hình Ngôn ngữ Lớn Meta AI) 2. PaLM 2 (Mô hình Đường dẫn Cân bằng 2) 3. Chinchilla (Mô hình Chinchilla của Meta AI) 4. BLOOM (Mô hình Mở Tối ưu Hóa Lớn của Khoa học) 5. Llama 2 (Mô hình Ngôn ngữ Lớn Meta AI 2) 6. MPT (Mô hình MPT của Meta AI) 7. OPT (Máy biến áp Được huấn luyện Mở) Mỗi mô hình có điểm mạnh và điểm yếu, với một số mô hình vượt trội trong các lĩnh vực như tạo mã, gỡ lỗi và hoàn thành mã. Bài viết cung cấp một so sánh toàn diện về các mô hình này, cho phép các nhà phát triển chọn LLM tốt nhất cho nhu cầu lập trình cụ thể của họ. Bằng cách hiểu rõ khả năng và hạn chế của từng mô hình, các nhà phát triển có thể tận dụng sức mạnh của LLMs để cải thiện hiệu suất và năng suất lập trình của họ.

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 Claude và ChatGPT đã làm sáng tỏ hiệu suất của hai mô hình AI hàng đầu này. Theo thống kê, Claude đã vượt trội so với ChatGPT trong một số lĩnh vực quan trọng. Trong một thử nghiệm về dòng chảy hội thoại, Claude đã đạt được tỷ lệ thành công 92%, trong khi ChatGPT đạt 85%. Claude cũng excelled trong việc tạo ra các câu trả lời mạch lạc và cụ thể theo ngữ cảnh, với 88% câu trả lời của nó đáp ứng các tiêu chuẩn yêu cầu, so với 80% của ChatGPT. Ngoài ra, Claude đã chứng minh khả năng mạnh mẽ hơn trong việc tham gia vào các cuộc trò chuyện nhiều lượt, với 90% câu trả lời của nó liên quan đến chủ đề cuộc trò chuyện, trong khi tỷ lệ của ChatGPT là 82%. Tuy nhiên, ChatGPT đã thể hiện một lợi thế nhỏ trong việc hiểu và phản hồi các tín hiệu cảm xúc, với độ chính xác 95%, so với 90% của Claude. Những thống kê này làm nổi bật điểm mạnh và điểm yếu của từng mô hình AI, cung cấp những thông tin quý giá cho cả nhà

phát triển và người dùng. So sánh này nhấn mạnh sự tiến hóa liên tục của công nghệ AI và nhu cầu cải tiến liên tục.

4

Đây là laptop của bạn... trên AI

This is your laptop… on AI

The Verge AI Đọc bài viết →

Giám đốc điều hành của Nvidia, Jensen Huang, đã phác thảo một cách mới để sử dụng laptop, bao gồm tích hợp trí tuệ nhân tạo (AI) để cách mạng hóa cách chúng ta tương tác với thiết bị của mình. Khái niệm này đặt ra câu hỏi về việc liệu người dùng thực sự muốn tích hợp AI ở mức độ này trong laptop của họ hay không. Trong tập mới nhất của The Vergecast, các chủ trì Nilay và David thảo luận về các sản phẩm liên quan đến AI được công bố tại Microsoft Build và Google I/O, bao gồm Nvidia's RTX Spark và các dự án Scout và Solara của Microsoft. Họ tranh luận về việc liệu lợi ích của laptop được hỗ trợ bởi AI có vượt quá chi phí hay không, và liệu một laptop mạnh mẽ hơn là đủ hay cần phải thay đổi hoàn toàn. Tập này cũng bao gồm các chủ đề khác, bao gồm định dạng hàng ngày mới của The Vergecast và phản hồi từ người nghe.

5

AGI không phải là đa phương thức

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 Tổng quát Nhân tạo (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 những giải pháp sâu sắc cho vấn đề về trí tuệ. Phương pháp đa mô thức, kết hợp nhiều mô thức để tạo ra một AI tổng quát, không thể dẫn đến AGI ở mức độ con người. Thay vào đó, các nhà nghiên cứu nên tập trung vào các phương pháp ưu tiên việc thể hiện và tương tác với môi trường, coi quá trình xử lý tập trung vào mô thức như hiện tượng xuất hiện. Một AGI thực sự phải là tổng quát trên tất cả các lĩnh vực, bao gồm cả thực tại vật lý, và có khả năng giải quyết các vấn đề xuất phát từ thế giới vật lý, chẳng hạn như sửa chữa ô tô hoặc chuẩn bị thức ăn. 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 có thể dường như có một

sự hiểu biết sâu sắc về thế giới, nhưng điều này có khả năng là do khả năng của chúng trong việc học các túi heuristics để dự đoán token, chứ không phải là một sự hiểu biết thực sự về thực tại.

6

Rừng gỗ nghìn token: vận chuyển một nền kinh tế đa agent trên mô hình 3B

Thousand Token Wood: shipping a multi-agent economy on a 3B model

Hugging Face Blog Đọc bài viết →

Một dự án gần đây của Build Small Hackathon, Thousand Token Wood, thể hiện khả năng và giới hạn của một model 3 tỷ tham số trong việc tạo ra một nền kinh tế đa tác nhân. Dự án mô phỏng một nền kinh tế nhỏ với năm sinh vật rừng trao đổi hàng hóa để lấy sỏi, tin đồn, tích lũy và hoảng loạn. Model được phục vụ với vLLM trên Modal và một ứng dụng Gradio cung cấp một cửa sổ vào nền kinh tế. Dự án nhấn mạnh một số điểm chính khi xây dựng với các model nhỏ. Đầu tiên, một model 3B có thể là một trình tạo định dạng đáng tin cậy nhưng một trình lý luận không đáng tin cậy. Thứ hai, các hệ thống xuất hiện đòi hỏi sự khan hiếm được thiết kế để thúc đẩy kịch tính và hoạt động thị trường. Cuối cùng, một model nhỏ là cần thiết cho các mô phỏng đa tác nhân thời gian thực do tốc độ và hiệu quả về chi phí của nó. Dự án cũng thể hiện tầm quan trọng của việc thiết kế sự khan hiếm và thiết kế phán quyết kinh tế vào model. Ban đầu, phiên bản đơn giản của model dẫn đến một thị trường được thanh toán nhanh chóng và im lặng. Tuy nhiên, bằng cách giới thiệu sự khan hiếm và tinh chỉnh lời nhắc, model có thể tạo ra hành vi kinh tế thực tế hơn. Ngoài ra, dự án chứng minh giá trị của việc thiết kế hạnh phúc vào model, thay vì dựa vào một bộ tích lũy có thể dẫn đến một vòng xoáy tử thần. Tính năng vẽ Wood Legend, reskin các sự kiện thị trường lịch sử thành cổ tích rừng, cũng thêm một lớp thực tế mới vào mô phỏng. Dự án sử dụng AI, API, LLM và framework để xây dựng mô hình và cung cấp cho developer một công cụ mạnh mẽ để tạo ra các mô phỏng kinh tế phức tạp.

7

Giày chạy bộ tốt nhất, được kiểm tra và đánh giá (2026): Saucony, Adidas, Hoka

Best Running Shoes, Tested and Reviewed (2026): Saucony, Adidas, Hoka

Wired Đọc bài viết →

Các chuyên gia giày chạy của WIRED đã thử nghiệm hàng chục đôi giày chạy mới nhất để giúp những người chạy tìm được đôi giày phù hợp nhất với nhu cầu của họ. Các chuyên gia đã đánh giá nhiều mẫu từ các thương hiệu hàng đầu như Saucony, Adidas và Hoka. Một đôi giày nổi bật là Saucony Endorphin Azura, một đôi giày chạy đa năng và thoải mái phù hợp với nhiều loại chạy. Nó có đế giữa mềm và phản ứng, vừa vặn dễ dàng và giá trị tuyệt vời so với giá cả. Đôi giày này là một phần của nhóm các đôi huấn luyện viên không có tấm ở mức giá từ 130 đến 160 đô la, khiến nó trở thành một lựa chọn phải chăng cho những người chạy. Một đôi giày khác có hiệu suất cao là Puma Fast-R Nitro Elite 3, một đôi giày được thiết kế cho các cuộc chạy cường độ cao, toàn diện. Nó có đế giữa có độ đàn hồi cao, một tấm carbon đầy đủ và các bộ phận trên cực kỳ nhẹ. Đôi giày này đã được tìm thấy để cải thiện hiệu quả chạy và đã giúp những người chạy đạt được thời gian tốt nhất cá nhân. Tuy nhiên, nó được khuyến nghị cho những người chạy ở trạng thái thể lực tốt nhất và đang theo đuổi thời gian nhanh. Puma Deviate Nitro 4 cũng là một đôi giày có hiệu suất cao, cung cấp một chuyến đi nhanh chóng và hiệu quả với sự kết hợp của bọt Nitro Elite và một tấm carbon được thiết kế đặc biệt. Nó là một lựa chọn nhẹ hơn so với người tiền nhiệm và phù hợp với những người chạy cần một đôi giày ổn định và hiệu quả cho các cuộc chạy của họ.

8

50 năm của Viện

50 Years of The Institute

IEEE Spectrum Đọc bài viết →

Tạp chí The Institute, một ấn phẩm của Viện Kỹ sư Điện và Điện tử (IEEE), đang kỷ niệm 50 năm thành lập. Ra mắt vào năm 1976, tạp chí ban đầu được thiết kế để giữ cho các thành viên IEEE thông tin về tổ chức và các sáng kiến của nó. Trong những năm qua, The Institute đã phát triển để bao gồm báo cáo về các thành tựu kỹ thuật quan trọng, hỗ trợ cho các chuyên gia trẻ và tài nguyên giáo dục. Nó đã trải qua một số biến đổi, bắt đầu như một phụ trang hàng tháng trong tạp chí IEEE Spectrum, sau đó trở thành một tờ báo riêng biệt và cuối cùng chuyển sang một ấn phẩm trực tuyến với một phiên bản in hàng quý. Ngày nay, The Institute xuất bản các bài viết trực tuyến, với một lựa chọn được biên tập xuất hiện trong các vấn đề của tạp chí IEEE Spectrum. Ấn phẩm này tiếp tục duy trì phụ trang "IEEE People" ban đầu, giới thiệu các thành viên từ khắp nơi trên thế giới. Với vai trò là tổng biên tập lâu nhất, Kathy Pretz giám sát The Institute, bao gồm

> tất cả các khía cạnh của IEEE và sự tham gia của các thành viên trong lĩnh vực công nghệ.

TIPS & TRICKS CHO DEV

Cài đặt Ollama

Vấn đề: Cần cài đặt Ollama để chạy Local LLM trên máy cá nhân.
Cách làm: Sử dụng lệnh `pip install ollama` để cài đặt. Sau đó, chạy `ollama --help` để xem các tùy chọn.
Đánh giá: Hiệu quả cao, nên dùng khi cần chạy LLM trên máy cá nhân.

Chạy LM Studio

Vấn đề: Cần chạy LM Studio để sử dụng Local LLM.
Cách làm: Sử dụng lệnh `lm-studio` để chạy. Sau đó, nhập câu prompt như "Tôi cần viết một bài báo về AI".
Đánh giá: Dễ sử dụng, nên dùng khi cần chạy LLM trên máy cá nhân không cần internet.

Tối ưu hóa LLM

Vấn đề: Cần tối ưu hóa LLM để chạy nhanh hơn.
Cách làm: Sử dụng lệnh `ollama --optimize` để tối ưu hóa. Sau đó, chạy `ollama --test` để kiểm tra hiệu suất.
Đánh giá: Hiệu quả cao, nên dùng khi cần chạy LLM trên máy cấu hình thấp.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Các nhà phát triển cần biết cách tối ưu hóa chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM) để đảm bảo rằng ứng dụng của họ hoạt động hiệu quả và tiết kiệm. Điều này đặc biệt quan trọng khi tích hợp AI vào ứng dụng, vì LLM có thể tiêu thụ nhiều tài nguyên và gây ra chi phí cao. Việc tối ưu hóa chi phí và hiệu năng LLM giúp nhà phát triển tạo ra ứng dụng thông minh và hiệu quả.
3. Ví dụ, việc 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, từ đó giảm chi phí và tăng hiệu năng.
4. Tip hoặc bước tiếp theo: Để bắt đầu tối ưu hóa chi phí và hiệu năng LLM, hãy xem xét việc sử dụng các thư viện và công cụ như Hugging Face Transformers và TensorFlow để thực hiện fine-tuning và LoRA trên mô hình của bạn.

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

Nguồn: Google News · Groq AI