BIS - MT · 05/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

> *"Muốn sang thì bắc cầu kiều, muốn con hay chữ thì yêu lấy thầy."*
>
> — Ca dao Việt Nam

Tôn trọng thầy cô, người dạy dỗ mình là nền tảng của việc học thành công.

TIN TỨC NỔI BẬT

1 Claude Code đấu với Cursor 2026: 80.8% SWE-bench, 1M Context [Đã thử nghiệm]

Claude Code vs Cursor 2026: 80.8% SWE-bench, 1M Context [Tested]

tech-insider.org Đọc bài viết →

Trong một bài kiểm tra hiệu suất gần đây, Claude Code và Cursor 2026 đã được so sánh để xác định khả năng của chúng. Bài kiểm tra, tập trung vào SWE-bench, một điểm chuẩn rộng rãi để đánh giá các model ngôn ngữ, cho thấy Claude Code đạt được điểm số ấn tượng 80,8%. Đây là một thành tựu đáng chú ý, vì nó chứng tỏ khả năng của model trong việc xử lý và hiểu chính xác các đầu vào ngôn ngữ phức tạp. Ngoài ra, bài kiểm tra đã đẩy các model đến giới hạn của chúng bằng cách cung cấp cho chúng một ngữ cảnh khổng lồ với 1 triệu token. Mặc dù trong kịch bản đầy thách thức này, Claude Code vẫn có thể hoạt động tốt, thể hiện sự mạnh mẽ và linh hoạt của nó. Kết quả của bài kiểm tra này cung cấp những thông tin quý giá về khả năng của Claude Code và tiềm năng ứng dụng của nó trong nhiều lĩnh vực, bao gồm xử lý ngôn ngữ tự nhiên và trí tuệ nhân tạo (AI). Kết quả kiểm tra này có thể sẽ thu hút sự quan tâm của các nhà phát triển (developer), nhà nghiên cứu và bất kỳ ai muốn tận dụng sức mạnh của các model ngôn ngữ (LLM) trong công việc của họ, đặc biệt là khi tích hợp với API hoặc xây dựng trên framework cụ thể.

2 Tình hình các Agent Framework: Chọn Runtime phù hợp cho Enterprise AI Execution

State of Agent Frameworks: Choosing the Right Runtime for Enterprise AI Execution

Medium Đọc bài viết →

Bài viết "Trạng thái của các Framework Trình điều khiển: Chọn Thời gian chạy Đúng cho Thực thi AI Doanh nghiệp" khám phá cảnh quan hiện tại của các framework trình điều khiển trong bối cảnh thực thi AI doanh nghiệp. Các framework này đóng vai trò là xương sống cho việc xây dựng, triển khai và quản lý các model AI, cho phép các tổ chức tích hợp AI vào hoạt động của họ. Bài viết nhấn mạnh tầm quan trọng của việc chọn framework trình điều khiển phù hợp với nhu cầu cụ thể của một tổ chức. Nó thảo luận về các yếu tố cần xem xét, chẳng hạn như khả năng mở rộng, linh hoạt và khả năng tích hợp. Các framework trình điều khiển khác nhau được so sánh, bao gồm Rasa, Microsoft Bot Framework và Google Cloud Dialogflow. Bài viết nhấn mạnh nhu cầu về một framework có thể tích hợp liền mạch với các hệ thống và cơ sở hạ tầng hiện có, đồng thời cung cấp các công cụ và tài nguyên cần thiết cho các developer xây dựng và triển khai các model AI một cách hiệu quả. Bằng cách chọn framework trình điều khiển phù hợp, các tổ chức có thể mở khóa toàn bộ tiềm năng của AI và cải thiện hiệu suất và khả năng cạnh tranh tổng thể. Bài viết nhằm cung cấp cái nhìn tổng quan toàn diện về trạng thái hiện tại của các framework trình điều khiển, giúp người đọc đưa ra quyết định sáng suốt khi chọn thời gian chạy cho thực thi AI doanh nghiệp của họ.

3

Xây dựng agent phân tích tài chính thông minh với LangGraph và Strands Agents

Build an intelligent financial analysis agent with LangGraph and Strands Agents

Amazon Web Services (AWS) Đọc bài viết →

Amazon Web Services (AWS) đã giới thiệu một giải pháp để xây dựng một tác nhân phân tích tài chính thông minh sử dụng LangGraph và Strands Agents. Cách tiếp cận đổi mới này cho phép tạo ra một công cụ phân tích tài chính tinh vi có thể xử lý và phân tích lượng lớn dữ liệu tài chính. LangGraph là một thư viện xử lý ngôn ngữ tự nhiên (NLP) cho phép các developer xây dựng và đào tạo các model học máy (machine learning) để hiểu và diễn giải dữ liệu tài chính. Strands Agents, mặt khác, là một framework phần mềm cho phép tạo ra các tác nhân thông minh có thể thực hiện các nhiệm vụ phức tạp, chẳng hạn như phân tích tài chính. Bằng cách kết hợp LangGraph và Strands Agents, các developer có thể xây dựng một tác nhân phân tích tài chính có thể trích xuất thông tin chi tiết từ dữ liệu tài chính, xác định xu hướng và mẫu, và cung cấp các khuyến nghị có thể thực hiện. Tác nhân thông minh này có thể được tích hợp với các hệ thống và công cụ

tài chính khác nhau, cho phép các doanh nghiệp đưa ra quyết định dựa trên dữ liệu và cải thiện hiệu suất tài chính của họ. Giải pháp này được thiết kế để giúp các doanh nghiệp tối ưu hóa quy trình phân tích tài chính của họ, giảm chi phí và cải thiện tình hình tài chính chung.

4

Tại sao LLM nên ngừng "nghĩ lớn tiếng" (và điều gì đến sau chain-of-thought)

Why LLMs should stop thinking out loud (and what comes after chain-of-thought)

BD Tech Talks Đọc bài viết →

Ngành công nghệ đang đối mặt với một tình huống khó khăn về đầu tư do chi phí cho token AI tăng vọt. Các gã khổng lồ công nghệ như Uber, Meta và Amazon đang đánh giá lại các quy trình AI của họ và áp đặt giới hạn cho chi tiêu tính toán AI nội bộ. Điều này chủ yếu là do chi phí cao để đào tạo các mô hình ngôn ngữ lớn (LLM) bằng cách sử dụng kỹ thuật Chain-of-Thought (CoT), bao gồm việc hướng dẫn LLM "nghĩ từng bước" trước khi đưa ra câu trả lời cuối cùng. Mặc dù CoT ban đầu là một giải pháp thông minh, cộng đồng AI đã rơi vào tâm lý cargo cult, đồng nhất việc tạo ra các token văn bản trung gian với quá trình xử lý nhận thức thực sự. Tuy nhiên, nghiên cứu đã chỉ ra rằng CoT là sự bắt chước của quá trình lý luận, chứ không phải là cơ chế thực sự của nó, và các token trung gian thường hoạt động như một ràng buộc cấu trúc chứ không phải là một sự khái quát hóa thuật toán thực sự. Điều này đã dẫn đến một nút thắt kỹ thuật và tài chính lớn, làm chậm tốc độ suy luận và che giấu bản chất thực sự của tính toán máy. Để mở rộng AI một cách bền vững, ngành công nghệ cần phải vượt qua CoT và khám phá các cơ chế lý luận thay thế.

5

Xây dựng bản đồ thế giới chỉ với 500 byte

Building a World Map with only 500 bytes

Simon Willison Đọc bài viết →

Một nhà phát triển, Iwo Kadziela, đã thành công trong việc tạo ra một bản đồ thế giới ASCII đáng tin cậy chỉ bằng 445 byte dữ liệu. Thành tựu này được thực hiện bằng cách sử dụng thuật toán nén dữ liệu deflate, một thuật toán nén dữ liệu. Dữ liệu nén sau đó được kết hợp bằng một đoạn mã JavaScript. Quá trình cũng liên quan đến việc sử dụng hàm fetch() với URI dữ liệu, một tính năng cho phép tải dữ liệu trực tiếp vào trình duyệt mà không cần yêu cầu máy chủ. Cách tiếp

cận sáng tạo này chứng tỏ tiềm năng cho việc đại diện và xử lý dữ liệu hiệu quả trong phát triển web. Bản đồ thế giới ASCII, được tạo ra thông qua phương pháp này, cung cấp một biểu diễn□□ và chức năng của địa lý thế giới.

6 Quảng cáo mới của Google hình dung Tuyên ngôn Độc lập được viết với sự trợ giúp của AI

New Google commercial imagines a Declaration of Independence written with help from AI

TechCrunch AI Đọc bài viết →

Google đã phát hành một đoạn phim thương mại tưởng tượng cách việc ký Tuyên ngôn Độc lập có thể trông như thế nào nếu các Founding Fathers có quyền truy cập vào Google Workspace. Đoạn quảng cáo, được đặt vào năm 1776, mô tả Thomas Jefferson và những người ký tên khác sử dụng các công cụ của Google, bao gồm Google Docs, Google Calendar và Google Meet, với sự giúp đỡ của AI. Những người sáng lập hư cấu sử dụng công cụ AI "help me visualize" của Google để thiết kế con dấu quốc gia và hỏi chatbot về lời khuyên. Đoạn quảng cáo là một cách nhìn hài hước về việc hợp tác với AI vào thế kỷ 18. Mặc dù đoạn quảng cáo đã nhận được hầu hết các bình luận tích cực trên YouTube và Instagram, một số nhà phê bình trên Bluesky đã gọi nó là "cringey" và "tone deaf", nhắm vào góc độ AI không thực tế.

7 T-Mobile di chuyển hàng chục nghìn virtual machine khỏi VMware giữa vụ kiện

T-Mobile moving tens of thousands of virtual machines off VMware amid lawsuit

Ars Technica Đọc bài viết →

T-Mobile đang tham gia vào một vụ kiện với Broadcom liên quan đến việc hỗ trợ các giấy phép vĩnh viễn của VMware. Công ty di động này tuyên bố rằng Broadcom, sau khi mua lại VMware, đã ngừng bán giấy phép vĩnh viễn để ủng hộ các dịch vụ đăng ký và sản phẩm đóng gói. T-Mobile đã mua giấy phép vĩnh viễn với hai năm hỗ trợ vào năm 2023, nhưng Broadcom từ chối gia hạn hỗ trợ cho năm thứ ba. Công ty này có hàng chục nghìn máy ảo sử dụng phần mềm VMware trên khoảng 303.140 lõi CPU và đang di chuyển khỏi nền tảng này do những thách thức tốn thời gian và kỹ thuật liên quan. T-Mobile đang tìm kiếm một phán quyết của tòa án rằng Broadcom có nghĩa vụ hợp đồng phải tiếp

tục hỗ trợ các giấy phép vĩnh viễn của VMware. Một thẩm phán đã cấp cho công ty này một lệnh cấm cho phép họ nhận được dịch vụ hỗ trợ cho đến tháng 8 năm 2026. Vụ việc này tương tự như một vụ việc đã được giải quyết riêng với AT&T và một vụ việc đang diễn ra với Tesco.

8

Cộng đồng fanfiction đang "chiến tranh" với AI — và chính họ

The fanfiction community is at war with AI — and itself

The Verge AI Đọc bài viết →

Cộng đồng fanfiction hiện đang tham gia vào một cuộc tranh luận nóng về việc sử dụng các công cụ AI tạo sinh, chẳng hạn như Claude và ChatGPT. Một phong trào mới đã xuất hiện, với một số người đọc và người viết cố gắng phát hiện và loại bỏ các tác giả sử dụng AI để viết fanworks. Một giao diện người dùng đã được phát triển cho kho lưu trữ fanfic phổ biến Archive of Our Own (AO3) mà □□ được cho là xác định các mã hóa còn lại bởi Claude. Giao diện người dùng này sẽ chuyển màu nền thành đỏ nếu nó phát hiện sự hiện diện của mã được Claude tiêm. Tuy nhiên, các phương pháp phát hiện đang được thực hiện là đáng ngờ, và có nguy cơ có các kết quả dương tính giả và khái quát hóa quá mức. Người tạo ra giao diện người dùng này nhằm mục đích bảo tồn yếu tố con người và tia sáng sáng tạo trong fandom, nhưng cộng đồng đã nhanh chóng tập hợp để công khai xấu hổ các nhà văn mà tác phẩm của họ được đánh dấu bởi công cụ này. Việc sử dụng AI trong các cộng đồng sáng tạo vẫn là một vấn đề gây tranh cãi, và cộng đồng fanfiction đang vật lộn để tìm sự cân bằng giữa đổi mới và tính xác thực.

TIPS & TRICKS CHO DEV

Tự động hóa test case

Vấn đề: Việc viết test case thủ công tốn thời gian và dễ xảy ra lỗi.

Cách làm: Sử dụng AI tools như TestComplete để tự động hóa test case, với lệnh `tsc --generate-test` để tạo test tự động. Ví dụ prompt "Generate test case for login function".

Đánh giá: Hiệu quả cao, giúp giảm thiểu thời gian viết test case, nhưng cần kiểm tra kết quả tự động hóa.

Code review tự động

Vấn đề: Việc review code thủ công tốn thời gian và dễ bỏ sót lỗi.

Cách làm: Sử dụng AI tools như SonarQube với lệnh `sonar-scanner` để review code tự động, với cấu hình `sonar.java.binaries`.
Đánh giá: Giúp phát hiện lỗi và cải thiện chất lượng code, nhưng cần thiết lập cấu hình phù hợp.

QA automation với API

Vấn đề: Việc kiểm tra API thủ công tốn thời gian và dễ xảy ra lỗi.
Cách làm: Sử dụng AI tools như Postman với lệnh `newman run` để kiểm tra API tự động, với ví dụ prompt "Run API test for user endpoint".
Đánh giá: Hiệu quả cao, giúp giảm thiểu thời gian kiểm tra API, nhưng cần thiết lập môi trường phù hợp.

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 hiểu cách tinh chỉnh và tối ưu hóa mô hình cho từng trường hợp sử dụng cụ thể. Điều này giúp giảm thiểu chi phí tính toán và thời gian phản hồi, đồng thời cải thiện hiệu suất của mô hình. Các kỹ thuật như fine-tuning và LoRA cho phép điều chỉnh mô hình để phù hợp với nhu cầu cụ thể.
3. Ví dụ, với mô hình LLaMA, chúng ta có thể sử dụng kỹ thuật LoRA để tinh chỉnh mô hình cho nhiệm vụ cụ thể như trả lời câu hỏi hoặc tạo văn bản.
4. Tip hoặc bước tiếp theo: Áp dụng kỹ thuật fine-tuning và LoRA để tối ưu hóa mô hình LLM cho ứng dụng của bạn, và theo dõi hiệu suất của mô hình để đảm bảo rằng nó đáp ứng được nhu cầu của người dùng.