BIS - MT · 30/05/2026

🤖 Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

✨ *"Challenges are what make life interesting; overcoming them is what makes life meaningful."*

↪ Thử thách là thứ làm cho cuộc sống thú vị; vượt qua chúng là thứ làm cho cuộc sống có ý nghĩa.

— Joshua J. Marine

💡 *Đừng sợ thử thách — chính những khó khăn ta vượt qua mới là những kỷ niệm đáng giá và nguồn tự hào thật sự.*

📰 TIN TỨC NỔI BẬT

1

Hướng dẫn hoàn chỉnh về tệp nhớ của AI Agent (CLAUDE.md, AGENTS.md, và hơn thế nữa)

🇬🇧 *The Complete Guide to AI Agent Memory Files (CLAUDE.md, AGENTS.md, and Beyond)*

📰 HackerNoon 🔗 Đọc bài viết →

> Bài viết "Hướng dẫn hoàn chỉnh về tệp nhớ của Trình đại diện AI" cung cấp một cái nhìn sâu sắc về các tệp nhớ được sử dụng bởi các trình đại diện AI. Những tệp này, bao gồm CLAUDE.md và AGENTS.md, lưu trữ thông tin quan trọng về kiến thức và hành vi của trình đại diện. Hướng dẫn giải thích mục đích và cấu trúc của các tệp này, cũng như vai trò của chúng trong việc định hình quá trình ra quyết định của trình đại diện. CLAUDE.md là một tệp siêu dữ liệu chứa thông tin về khả năng của trình đại diện, chẳng hạn như khả năng hiểu và tạo ngôn ngữ của nó. AGENTS.md, mặt khác, là một tệp cấu hình định nghĩa hành vi của trình đại diện và tương tác với môi trường của nó. Bài viết cũng đề cập đến các tệp khác liên quan, chẳng hạn như chính sách và model, đóng góp vào hiệu suất tổng thể của trình đại diện. Hiểu biết về các tệp nhớ này là điều cần thiết cho các nhà phát triển làm việc với các trình đại diện AI, vì nó cho phép họ tùy chỉnh và tối ưu hóa hành vi của trình đại diện để phù hợp với các trường hợp sử dụng cụ thể. Hướng dẫn nhằm cung cấp một cái nhìn tổng quan toàn diện về chủ đề, khiến nó trở thành một tài nguyên quý giá cho bất kỳ ai quan tâm đến việc phát triển trình đại diện AI.

2 Lớp lệnh AI nhà kho đa agent cho phép xuất sắc hoạt động và thông minh chuỗi cung ứng

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

📰 NVIDIA Developer 🔗 Đọc bài viết →

> NVIDIA đã giới thiệu Lớp lệnh AI Nhà kho đa tác nhân, một công nghệ tiên tiến được thiết kế để tối ưu hóa hoạt động nhà kho và quản lý chuỗi cung ứng. Giải pháp đổi mới này tận dụng trí tuệ nhân tạo (AI) để tự động hóa các quy trình, tăng cường hiệu quả và cung cấp thông tin theo thời gian thực về các hoạt động trong nhà kho. Lớp lệnh AI Nhà kho đa tác nhân cho phép giao tiếp và phối hợp liền mạch giữa các hệ thống nhà kho khác nhau, bao gồm robot, xe nâng và các thiết bị khác. Bằng cách tích hợp việc ra quyết định dựa trên AI, hệ thống có thể tối ưu hóa các nhiệm vụ như quản lý hàng tồn kho, thực hiện đơn hàng và hậu cần. Công nghệ này cũng cung cấp thông tin chuỗi cung ứng, cho phép các doanh nghiệp đưa ra quyết định dựa trên dữ liệu và phản ứng nhanh chóng với các thay đổi về nhu cầu hoặc cung ứng. Lớp lệnh AI Nhà kho đa tác nhân được xây dựng trên nền tảng tính toán AI của NVIDIA, cung cấp khả năng xử lý và khả năng mở rộng cần thiết để hỗ trợ các khối lượng công việc AI phức tạp. Bằng cách triển khai giải pháp này, các doanh nghiệp có thể đạt được sự xuất sắc về hoạt động, giảm chi phí và cải thiện sự hài lòng của khách hàng. Công nghệ này có tiềm năng cách mạng hóa cách thức hoạt động của các nhà kho, khiến chúng trở nên hiệu quả, linh hoạt và phản ứng nhanh với các điều kiện thị trường thay đổi.

3 Nhà máy agent: Kết nối agent, ứng dụng và dữ liệu với các tiêu chuẩn mở mới như MCP và A2A

🇬🇧 *Agent Factory: Connecting agents, apps, and data with new open standards like MCP and A2A*

📰 Microsoft Azure 🔗 Đọc bài viết →

> Microsoft đã giới thiệu Agent Factory, một nền tảng mới nhằm kết nối các tác nhân, ứng dụng và dữ liệu thông qua các tiêu chuẩn mở. Nền tảng này sử dụng Cloud PC (MCP) và tiêu chuẩn Application-to-Application (A2A) của Microsoft để tạo điều kiện cho sự tương tác liền mạch giữa các thành phần khác nhau. Agent Factory cho phép các nhà phát triển tạo và triển khai các tác nhân thông minh có thể tương tác với các ứng dụng và nguồn dữ liệu khác nhau. Các tác nhân này có thể được sử dụng để tự động hóa các nhiệm vụ, nâng cao trải nghiệm

người dùng và cung cấp thông tin chi tiết thông qua phân tích dữ liệu. Việc giới thiệu các tiêu chuẩn MCP và A2A dự kiến sẽ đơn giản hóa quá trình phát triển và cải thiện khả năng tương tác giữa các hệ thống khác nhau. Điều này, đến lượt, có thể dẫn đến tăng hiệu suất và năng suất trong các ngành công nghiệp khác nhau. Bằng cách cung cấp một khuôn khổ tiêu chuẩn hóa cho việc phát triển và tương tác tác nhân, Agent Factory có tiềm năng cách mạng hóa cách các ứng dụng và dữ liệu được kết nối và sử dụng. Việc chuyển hướng của Microsoft towards các tiêu chuẩn mở dự kiến sẽ thúc đẩy đổi mới và hợp tác trong ngành công nghệ, đặc biệt là trong lĩnh vực AI, API, LLM, model, token, và framework.

4

Mạng botnet với hơn 17 triệu thiết bị bị phá vỡ

🇬🇧 *Botnet of more than 17 million devices dismantled*

📰 Ars Technica 🔗 Đọc bài viết →

Cơ quan chức năng Hà Lan, hợp tác với Trung tâm An ninh mạng Quốc gia, đã phá vỡ một botnet khổng lồ bao gồm hơn 17 triệu thiết bị. Hoạt động này được khởi xướng sau khi một nhà nghiên cứu an ninh báo cáo mạng lưới rộng lớn này cho cơ quan chức năng. Botnet này được liên kết với một mạng proxy cư trú của Nga, ASOCKS, cung cấp dịch vụ che giấu vị trí hoặc danh tính. Các dịch vụ này thường được sử dụng cho các mục đích bất hợp pháp như tấn công DDoS, hoạt động phishing và thu thập nội dung trang web. Cơ sở hạ tầng lưu trữ của botnet được đặt tại Hà Lan, và cơ quan chức năng đã thu giữ một số máy chủ để điều tra. Nhà cung cấp lưu trữ sau đó đã đưa botnet offline do sử dụng cho mục đích tội phạm. Sự việc này nhấn mạnh tầm quan trọng của việc cài đặt các bản cập nhật bảo mật một cách kịp thời và nghiên cứu cẩn thận các ứng dụng trước khi cài đặt để ngăn chặn thiết bị bị cuốn vào botnet.

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, những mô hình này đã xuất hiện không phải vì những giải pháp sâu sắc cho vấn đề trí tuệ, mà vì chúng đã được mở rộng hiệu quả trên

phần cứng hiện có. Phương pháp đa mô thức, liên quan đến việc kết hợp nhiều mô thức để tạo ra một AI tổng quát, được coi là con đường dẫn đến AGI. Tuy nhiên, chiến lược này không thể thành công trong ngắn hạn, vì nó không giải quyết được các không gian vấn đề quan trọng mà một AGI thực sự phải có thể giải quyết. Một AGI thực sự phải có thể giải quyết các vấn đề bắt nguồn từ thực tế vật lý, chẳng hạn như sửa chữa ô tô hoặc chuẩn bị thức ăn. Những vấn đề này đòi hỏi một hình thức trí tuệ cơ bản nằm trong một mô hình thế giới vật lý. Các Mô hình Ngôn ngữ Lớn (LLM) hiện tại đã được chứng minh là có khả năng trong các nhiệm vụ ngôn ngữ, nhưng sự hiểu biết của chúng về thế giới là nông cạn và dựa trên các quy tắc kinh nghiệm hơn là sự hiểu biết sâu sắc về thực tế. Ý tưởng rằng LLM học các mô hình ngầm của thế giới thông qua dự đoán token tiếp theo là một lý thuyết phổ biến, nhưng không rõ liệu điều này có thực sự xảy ra hay không. Mặc dù một số nghiên cứu cho thấy LLM có thể dự đoán trạng thái của một hệ thống vật lý, chẳng hạn như một bàn cờ Othello, nhưng điều này không nhất thiết chuyển thành sự hiểu biết sâu sắc về thế giới vật lý.

6

Tạo hồ sơ trong PyTorch (Phần 1): Hướng dẫn dành cho người mới bắt đầu về torch.profiler

🇬🇧 *Profiling in PyTorch (Part 1): A Beginner's Guide to torch.profiler*

📰 Hugging Face Blog 🔗 Đọc bài viết →

Hướng Dẫn Cơ Bản về PyTorch Profiling: torch.profiler Quá trình phân tích hiệu suất là một bước quan trọng trong việc tối ưu hóa các model PyTorch, cho dù đó là để cải thiện tốc độ suy luận hay hiểu tại sao các vòng lặp đào tạo lại chậm hơn dự kiến. Tuy nhiên, việc phân tích hiệu suất có thể gây khó khăn do sự phức tạp của các dấu vết phân tích. Hướng dẫn này dành cho người mới bắt đầu nhằm mục đích đơn giản hóa quá trình bằng cách hướng dẫn từng bước cách đọc các dấu vết phân tích và sử dụng chúng để thúc đẩy tối ưu hóa. Hướng dẫn bắt đầu với một kịch bản PyTorch cơ bản thực hiện các phép toán nhân ma trận và cộng. Nó sử dụng mô-đun torch.profiler để phân tích hiệu suất các phép toán và cung cấp lời giải thích chi tiết về bảng phân tích hiệu suất kết quả. Bảng này được chia thành các cột hiển thị các sự kiện được kích hoạt, thời gian thực hiện trên CPU hoặc GPU và số lượng cuộc gọi. Hướng dẫn cũng giới thiệu các khái niệm quan trọng, chẳng hạn như cách các phép toán PyTorch được dịch sang các nhân GPU và cách mô-đun torch.compile hoạt động. Vào cuối hướng dẫn, người đọc nên có thể hiểu được các kiến thức cơ bản về phân tích hiệu suất và cách sử dụng nó để tối ưu hóa các model PyTorch của mình. Hướng

dẫn được thiết kế để trở thành một tài liệu đọc dễ dàng với những khoảnh khắc "Aha!", giúp nó trở nên dễ tiếp cận với người mới bắt đầu.

7

(không có nội dung)

🇬🇧

📰 LangChain Blog 🔗 Đọc bài viết →

Một công cụ nghiên cứu mới được hỗ trợ bởi AI đã được phát triển để phân tích dữ liệu GDP trên 27 quốc gia thành viên EU. Công cụ này, được gọi là Finance Research API, có thể phát hiện các bất thường và xác định các yếu tố cấu trúc và chu kỳ đằng sau sự tăng trưởng hoặc thu hẹp kinh tế ở cấp độ ngành. Nó tạo ra một bản báo cáo 13 phần trong khoảng 45 phút, với chi phí khoảng 2,20 đô la cho các cuộc gọi API. API này kết hợp dữ liệu cấu trúc được cấp phép với thông tin trực tuyến, bao gồm cả bình luận của ngân hàng trung ương và phân tích cấp độ ngành. Trong một thử nghiệm, công cụ này đã xác định Ireland là quốc gia có độ lệch lớn nhất do sự tăng đột ngột về xuất khẩu dược phẩm, trong khi Đức được đánh dấu vì sự thu hẹp cấu trúc do tiếp xúc với ngành ô tô và sụp đổ xây dựng. Kiến trúc của công cụ này bảo tồn nhật ký quyết định, cho phép người dùng theo dõi bất kỳ điểm dữ liệu nào trở lại nguồn gốc của nó. Finance Research API có thể xử lý các truy vấn nghiên cứu phức tạp và trả về các câu trả lời dựa trên dữ liệu công khai và riêng tư, với trích dẫn trực tuyến. Nó được thiết kế để giúp các bộ phận nghiên cứu vĩ mô xác định các quốc gia hoạt động bất thường và hiểu các yếu tố cơ bản đằng sau sự tăng trưởng hoặc thu hẹp kinh tế.

8

Các công ty công nghệ tuyệt vọng muốn quay phim bạn làm việc nhà

🇬🇧 *Tech companies desperately want to film you doing chores*

📰 The Verge AI 🔗 Đọc bài viết →

Các công ty công nghệ đang tìm cách thu thập dữ liệu từ thế giới thực để đào tạo robot của họ, có thể thực hiện các công việc gia đình nhàm chán. Để đạt được điều này, các công ty đang trả tiền cho người dân để ghi lại các hoạt động hàng ngày của họ, chẳng hạn như dọn dẹp, nấu ăn và giặt là, để đổi lấy bồi thường. Ví dụ, công ty khởi nghiệp đào tạo AI Shift đang cung cấp dịch vụ dọn dẹp miễn phí cho người New

> York để đổi lấy cảnh quay của nhân viên dọn dẹp thực hiện các nhiệm vụ như chà rửa bát đĩa và lau sàn. Phương pháp này đang được sao chép bởi các công ty khác, bao gồm Pronto ở Ấn Độ, sử dụng nhà của khách hàng để thu thập cảnh quay đào tạo AI. Một số công ty khởi nghiệp cũng sử dụng nắp máy ảnh hoặc ứng dụng để thu thập dữ liệu từ góc nhìn thứ nhất từ người lao động tạm thời, trong khi những công ty khác đang tạo ra các trang trại dữ liệu nơi người lao động được trả tiền để thực hiện các nhiệm vụ lặp đi lặp lại. Mục tiêu là tạo ra tài liệu đào tạo AI có giá trị giúp các công ty phát triển robot có khả năng thực hiện các công việc gia đình, đây là một cơ hội béo bở trên thị trường.

⚡ TIPS & TRICKS CHO DEV

⚡ Tối ưu hóa Code

Vấn đề: Code không tối ưu dẫn đến hiệu suất kém.
Cách làm: Sử dụng GitHub Copilot để tự động đề xuất cải tiến. Ví dụ, nhập prompt "optimize this code" để nhận đề xuất.
Đánh giá: Tối ưu hóa code hiệu quả, nhưng cần kiểm tra kỹ trước khi áp dụng.

⚡ Tự động hoàn thiện

Vấn đề: Gõ code thủ công mất thời gian.
Cách làm: Sử dụng GitHub Copilot để tự động hoàn thiện code. Ví dụ, nhập prompt "complete this function" để nhận đề xuất.
Đánh giá: Tiết kiệm thời gian, nhưng cần kiểm tra độ chính xác.

⚡ Kiểm tra lỗi

Vấn đề: Khó tìm lỗi trong code.
Cách làm: Sử dụng GitHub Copilot để tìm kiếm và sửa lỗi. Ví dụ, nhập prompt "debug this code" để nhận đề xuất.
Đánh giá: Tìm lỗi hiệu quả, nhưng cần kiểm tra kỹ trước khi áp dụng.

📚 BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Dev cần biết cách tối ưu chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM) để đảm bảo ứng dụng AI của họ hoạt động hiệu quả và tiết kiệm. Điều này đặc biệt quan trọng khi tích hợp AI vào các ứng dụng thực tế. Việc tối ưu hóa có thể giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý.
3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) để điều chỉnh mô hình LLM cho phù hợp với use case cụ thể, từ đó giảm thiểu số

lượng tham số cần huấn luyện và tăng tốc độ xử lý.

4. 💡 Tip hoặc bước tiếp theo: Hãy xem xét việc sử dụng các thư viện như Hugging Face Transformers hoặc các công cụ tối ưu hóa như Optuna để tự động hóa quá trình tối ưu hóa mô hình LLM và đạt được hiệu suất tốt nhất cho ứng dụng của bạn.