BIS - MT · 30/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Nước chảy đá mòn."

— Tục ngữ Việt Nam

Sự kiên trì bền bỉ cuối cùng sẽ thắng mọi trở ngại — dù khó khăn đến đâu, nỗ lực liên tục sẽ tạo ra thay đổi.

TIN TỨC NỔI BẬT

1

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 người dùng tạo một công cụ phân tích tài chính tinh vi có thể xử lý và phân tích lượng lớn dữ liệu tài chính. LangGraph, một thư viện xử lý ngôn ngữ tự nhiên (NLP), được sử dụng để trích xuất thông tin liên quan từ các tài liệu tài chính, chẳng hạn như báo cáo và tuyên bố tài chính. Strands Agents, một nền tảng low-code, được sử dụng để xây dựng tác nhân phân tích tài chính, cho phép người dùng tạo một giải pháp tùy chỉnh mà không cần kiến thức mã hóa rộng rãi. Tác nhân phân tích tài chính thông minh có thể thực hiện các nhiệm vụ như trích xuất dữ liệu, nhận dạng thực thể và phân tích cảm xúc, cung cấp cho người dùng những thông tin quý giá về hiệu suất và xu hướng tài chính. Bằng cách tận dụng LangGraph và Strands Agents, người dùng có thể tự động hóa phân tích tài chính, giảm công sức thủ công và đưa ra quyết định sáng suốt hơn. Giải pháp này được thiết kế để có khả năng mở rộng và linh hoạt, phù hợp với các nguồn dữ liệu tài chính và yêu cầu phân tích khác nhau. Bằng cách tận dụng sức mạnh của LangGraph và Strands Agents, người dùng có thể xây dựng một tác nhân phân tích tài chính thông minh và mạnh mẽ, nâng cao khả năng ra quyết định tài chính của họ.

2 Các open-weight model tốt nhất của Trung Quốc — và những đối thủ mạnh nhất từ Mỹ

The best Chinese open-weight models — and the strongest US rivals

understandingai.org Đọc bài viết →

Bài viết thảo luận về các mô hình trọng lượng mở hàng đầu của Trung Quốc và các đối thủ mạnh nhất của Mỹ. Trong lĩnh vực trí tuệ nhân tạo, các mô hình trọng lượng mở đề cập đến các hệ thống AI có thể được đào tạo trên một loạt các nhiệm vụ mà không bị ràng buộc vào một nhiệm vụ hoặc tập dữ liệu cụ thể. Các công ty Trung Quốc đã đạt được những tiến bộ đáng kể trong lĩnh vực này, với một số mô hình nổi bật nhờ khả năng của chúng. Các mô hình Trung Quốc được đề cập trong bài viết bao gồm LLaMA, một mô hình ngôn ngữ lớn được phát triển bởi Meta, và Chengyu, một mô hình đa phương thức có thể xử lý cả văn bản và hình ảnh. Những mô hình này đã thể hiện hiệu suất ấn tượng trong các nhiệm vụ khác nhau, bao gồm dịch ngôn ngữ, tạo văn bản và nhận dạng hình ảnh. Bài viết cũng đề cập đến các đối thủ của Mỹ, bao gồm mô hình ngôn ngữ phổ biến Llama, được phát triển bởi Meta, cũng được biết đến với tên gọi Large Language Model Meta AI. Tuy nhiên, các mô hình cụ thể của Mỹ được đề cập không được chi tiết trong nội dung được cung cấp.

3 MiniMax-M2 là "vua" mới của các open source LLM (đặc biệt xuất sắc trong agentic tool calling)

MiniMax-M2 is the new king of open source LLMs (especially for agentic tool calling)

VentureBeat Đọc bài viết →

MiniMax-M2 đã nổi lên như một mô hình ngôn ngữ lớn (LLM) mã nguồn mở hàng đầu trong ngành công nghệ. Đặc biệt, nó excels trong việc gọi công cụ đại lý, một tính năng cho phép mô hình tương tác với các công cụ và dịch vụ bên ngoài một cách tự chủ hơn. Khả năng này cho phép MiniMax-M2 thực hiện các nhiệm vụ phức tạp đòi hỏi nhiều bước và tương tác với các hệ thống khác nhau. Mô hình MiniMax-M2 là mã nguồn mở, khiến nó có thể tiếp cận được với các nhà phát triển và nhà nghiên cứu có thể đóng góp vào sự phát triển và tùy chỉnh của nó. Sự cởi mở này thúc đẩy sự hợp tác và đổi mới, có thể dẫn đến những tiến bộ đáng kể trong lĩnh vực LLM. Mặc dù các chi tiết cụ thể về kiến trúc và hiệu suất của mô hình không được cung cấp, danh tiếng của nó như một LLM mã nguồn mở hàng đầu cho thấy nó đã thể hiện khả năng ấn tượng trong việc gọi công cụ đại lý và các lĩnh vực

khác. Khi ngành công nghệ tiếp tục phát triển, mô hình MiniMax-M2 có khả năng sẽ đóng vai trò quan trọng trong việc định hình tương lai của LLM và các ứng dụng của nó.

4 Tại sao block-sparse featurizer của Goodfire là một bước đột phá trong AI interpretability

Why Goodfire's block-sparse featururizers are a breakthrough in AI interpretability

BD Tech Talks Đọc bài viết →

Các nhà nghiên cứu tại Goodfire AI đã phát triển một kỹ thuật mới gọi là Block-Sparse Featurizers (BSF) để cải thiện khả năng giải thích của các mạng nơ-ron sâu. Không giống như các phương pháp hiện có như Sparse Autoencoders (SAEs), phân rã các hoạt động của model thành các hướng một chiều cách ly, BSF phân rã chúng thành các không gian con đa chiều. Cách tiếp cận này cung cấp một lời giải thích chi tiết hơn về các thành phần bên trong của model và có thể dẫn đến khả năng điều khiển và kiểm soát tốt hơn trong các ứng dụng AI. Tiêu chuẩn vàng hiện tại của tính giải thích cơ chế, SAEs, đã được chứng minh là có những hạn chế trong việc nắm bắt các khái niệm phức tạp trong các mạng nơ-ron, đặc biệt là trong các model tầm nhìn. Những model này tự nhiên nằm trên các không gian hình học liên tục, thấp chiều gọi là đa tạp, được SAEs chia thành các điểm cách ly, cứng nhắc. Phương pháp BSF của Goodfire giả định rằng một số nhỏ các khối đa chiều được kích hoạt tại bất kỳ thời điểm nào, cho phép một biểu diễn thống nhất và liên tục hơn về các khái niệm. Kỹ thuật mới này có tiềm năng mở ra con đường cho sự hiểu biết và kiểm soát tốt hơn các model AI phức tạp.

5 Cách một geothermal plant bị bỏ quên có được cơ hội thứ hai

How an overlooked geothermal plant got a second chance

MIT Tech Review Đọc bài viết →

Một công ty khởi nghiệp địa nhiệt nhỏ, Zanskar, đã thành công trong việc hồi sinh một nhà máy điện đang gặp khó khăn ở New Mexico, được gọi là Lightning Dock. Nhà máy này, vốn đang gặp khó khăn do nhiệt độ nước giảm, đã được Zanskar mua lại vào tháng 6 năm 2024. Sử dụng mô hình hóa tiên tiến và công nghệ khoan hiện đại, công ty đã xác định một vị trí giếng mới và khoan xuống hàng nghìn feet để tiếp cận một hồ chứa ngầm nóng hơn. Giếng mới, có độ sâu 8.000 feet,

đã tăng lưu lượng nước và nhiệt độ, cho phép nhà máy hoạt động với công suất đầy đủ một lần nữa. Thành tựu này có ý nghĩa quan trọng vì nó chứng tỏ tiềm năng của năng lượng địa nhiệt trong việc cung cấp điện không phát thải 24/7. Sự thành công của dự án cũng có thể có ý nghĩa đối với các địa điểm địa nhiệt khác, những nơi có thể được hưởng lợi từ việc khoan sâu hơn để tiếp cận các hồ chứa nóng hơn. Nhà máy Lightning Dock, có công suất 15 megawatt, đã tạo ra hơn gấp đôi lượng điện so với các giếng cũ, và dự kiến sẽ tiếp tục hoạt động hiệu quả trong nhiều năm tới.

6

OlmoEarth Platform: Geospatial inference ở planetary scale

The OlmoEarth Platform: Geospatial inference at planetary scale

Hugging Face Blog Đọc bài viết →

Nền tảng OlmoEarth là một hệ thống suy luận không gian địa lý được thiết kế để xử lý dữ liệu vệ tinh quy mô lớn. Nền tảng này được tạo ra để giải quyết các thách thức khi chạy các model học máy tại quy mô hành tinh, nơi dữ liệu đầu vào có thể bao gồm nhiều dải phổ, loại cảm biến và bước thời gian trên một khu vực địa lý lớn. Hệ thống phải xử lý sự cố tại quy mô, xử lý hình ảnh hiệu quả và ghép kết quả thành bản đồ nhất quán về mặt địa lý. Các model OlmoEarth được đào tạo trước trên 10 terabyte dữ liệu vệ tinh đa phương thức và đã được điều chỉnh cho các ứng dụng như giám sát phá rừng, an ninh lương thực và rủi ro cháy rừng. Cơ sở hạ tầng của nền tảng được thiết kế để chạy model một cách tiết kiệm chi phí tại đúng thời điểm và địa điểm, giám sát hiệu suất và chuyển đổi đầu ra thô thành thông tin có thể hành động. Nó có thể chạy suy luận trên các khu vực quy mô lục địa trong khoảng một ngày, xử lý hàng chục terabyte hình ảnh với chi phí chỉ một phần nhỏ của một xu mỗi kilômét vuông. Các thách thức kỹ thuật của hệ thống bao gồm tìm kiếm và truy cập hình ảnh vệ tinh, căn chỉnh nó trên các phép chiếu và độ phân giải, và xử lý nó một cách hiệu quả. Nền tảng OlmoEarth đã phát triển các giải pháp cho những thách thức này, bao gồm chia các công việc thành ba giai đoạn, mỗi giai đoạn được ghép với một hồ sơ phần cứng riêng biệt, và phân phối các giai đoạn này trên nhiều máy.

7

Trích lời D. Richard Hipp

Quoting D. Richard Hipp

Simon Willison Đọc bài viết →

Trong một tuyên bố gần đây, D. Richard Hipp lưu ý về tác động đáng kể của SQL đối với phong cảnh lập trình. Trước khi có sự ra đời của SQL, các cá nhân được gọi là lập trình viên COBOL chịu trách nhiệm tạo ra phần mềm để truy vấn các tập dữ liệu lớn. Nhiệm vụ này vừa tốn thời gian vừa tốn kém. Tuy nhiên, với sự giới thiệu của SQL, người dùng đã có khả năng chỉ định các truy vấn một cách đơn giản và tiện lợi, tự động hóa quá trình tạo ra mã cần thiết. Kết quả là, vai trò của các lập trình viên COBOL không biến mất, mà thay vào đó đã phát triển, khi các lập trình viên thích nghi với công nghệ mới và tập trung vào các khía cạnh khác của công việc. Sự thay đổi này nhấn mạnh sức mạnh chuyển đổi của SQL trong việc đơn giản hóa việc truy vấn và quản lý dữ liệu.

8

Mythos phát hiện các điểm yếu crypto đã tồn tại nhiều năm mà không ai biết

Mythos uncovers crypto weaknesses that went unknown for years

Ars Technica Đọc bài viết →

Một đột phá gần đây của mô hình bảo mật AI, Mythos, của Anthropic đã phát hiện ra những điểm yếu trong hai thuật toán mã hóa được sử dụng rộng rãi. Thuật toán đầu tiên, HAWK, là một sơ đồ chữ ký số chống lượng tử đang được xem xét như một tiêu chuẩn chính thức của Mỹ. Mặc dù đã chịu đựng được nhiều năm thử nghiệm, Mythos đã tìm thấy một điểm yếu khiến nó bị hỏng. Điểm yếu này có thể được giảm thiểu bằng cách tăng gấp đôi kích thước khóa, nhưng điều này sẽ làm cho HAWK ít mong muốn hơn so với các thuật toán khác có sẵn. Thuật toán thứ hai bị ảnh hưởng là AES, một mã hóa được sử dụng rộng rãi. Mythos đã tìm thấy một phương pháp để giảm vừa phải công việc cần thiết để đánh bại AES, nhưng điều này không làm hỏng hệ thống. Những phát hiện này rất quan trọng vì chúng có thể báo hiệu những tiến bộ quan trọng trong việc phá vỡ mã hóa quan trọng đối với quyền riêng tư và bảo mật. Tuy nhiên, các chuyên gia lưu ý rằng những cuộc tấn công này là dần dần và không làm hỏng bất kỳ hệ thống mã hóa nào được sử dụng ngày nay.

TIPS & TRICKS CHO DEV

Tạo Code Mới

Vấn đề: Tạo code mới từ scratch tốn thời gian.

Cách làm: Sử dụng Claude Code với lệnh `claude create` và cung cấp prompt như "Write a Python function to sort a list of numbers".
Đánh giá: Hiệu quả khi cần tạo code nhanh, nhưng nên kiểm tra kết quả.

Debug Code

Vấn đề: Debug code phức tạp mất nhiều thời gian.
Cách làm: Sử dụng lệnh `claude debug` và cung cấp thông tin về lỗi như "Debug this Python code: ...".
Đánh giá: Hiệu quả khi cần tìm lỗi nhanh, nhưng cần kiểm tra kết quả.

Refactor Code

Vấn đề: Refactor code để tối ưu hóa hiệu suất tốn thời gian.
Cách làm: Sử dụng lệnh `claude refactor` và cung cấp prompt như "Optimize this Python code for performance".
Đánh giá: Hiệu quả khi cần tối ưu hóa code, nhưng nên kiểm tra kết quả.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Dev cần biết về tối ưu chi phí và hiệu năng LLM để giảm thiểu chi phí vận hành và tăng tốc độ xử lý của mô hình AI. Điều này đặc biệt quan trọng khi triển khai ứng dụng AI trên quy mô lớn. Việc tối ưu hóa giúp giảm tải cho hệ thống và cải thiện trải nghiệm người dùng.
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 yêu cầu về tài nguyên và cải thiện hiệu suất.
4. Tip hoặc bước tiếp theo: Sử dụng các công cụ như Hugging Face Transformers để thực hiện fine-tuning và LoRA cho mô hình LLM của bạn, và đánh giá hiệu suất của mô hình trên các chỉ số như latency và throughput.