BIS - MT · 05/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Keep your eyes on the stars and your feet on the ground.”

↪ Hãy để mắt nhìn lên các vì sao và chân đặt vững trên mặt đất.

— Theodore Roosevelt

Cân bằng giữa tầm nhìn xa và hành động thực tiễn — mơ lớn nhưng cũng phải làm việc cần cù, thực tế để biến giấc mơ thành hiện thực.

TIN TỨC NỔI BẬT

1

AWS mã nguồn mở máy chủ MCP cho Bedrock AgentCore để đơn giản hóa phát triển agent AI

AWS Open-Sources an MCP Server for Bedrock AgentCore to Streamline AI Agent Development

MarkTechPost Đọc bài viết →

Amazon Web Services (AWS) đã mở mã nguồn một máy chủ MCP (Nền tảng Đa đám mây) cho Bedrock AgentCore, một framework được thiết kế để đơn giản hóa việc phát triển đại lý AI. Động thái này nhằm thúc đẩy sự hợp tác và đổi mới trong lĩnh vực trí tuệ nhân tạo. Máy chủ MCP là một thành phần quan trọng của Bedrock AgentCore, cho phép các nhà phát triển xây dựng, thử nghiệm và triển khai các đại lý AI trên nhiều nền tảng đám mây. Bằng cách mở mã nguồn máy chủ MCP, AWS cung cấp một framework tiêu chuẩn hóa và mở cho việc phát triển đại lý AI, cho phép các nhà phát triển tập trung vào việc tạo ra các đại lý thông minh mà không phải lo lắng về cơ sở hạ tầng cơ bản. Việc mở mã nguồn máy chủ MCP dự kiến sẽ đẩy nhanh việc phát triển các đại lý AI và thúc đẩy một cách tiếp cận dựa trên cộng đồng để đổi mới AI. Động thái này phù hợp với cam kết của AWS trong việc làm cho AI trở nên dễ tiếp cận và tiết kiệm hơn cho các nhà phát triển và tổ chức. Bằng cách đóng góp vào framework Bedrock AgentCore, AWS đang mở rộng thêm các dịch vụ của mình trong không gian AI và thể hiện sự cam kết của mình trong việc thúc đẩy đổi mới trong lĩnh vực này.

2

Xây dựng agent phân tích tài chính thông minh với LangGraph và Strands Agents | Amazon Web Services

Build an intelligent financial analysis agent with LangGraph and Strands Agents | Amazon Web Services

Amazon Web Services (AWS) Đọc bài viết →

> Amazon Web Services (AWS) đã giới thiệu một giải pháp mới để xây dựng các tác nhân phân tích tài chính thông minh bằng cách sử dụng LangGraph và Strands Agents. Cách tiếp cận đổi mới này cho phép người dùng tạo ra các công cụ phân tích tài chính tinh vi có thể xử lý và phân tích lượng lớn dữ liệu tài chính. LangGraph là một mô hình ngôn ngữ dựa trên đồ thị có thể hiểu các khái niệm và mối quan hệ tài chính phức tạp, cho phép nó trích xuất thông tin liên quan từ tài liệu và dữ liệu tài chính. Strands Agents, mặt khác, là một nền tảng low-code cho phép người dùng xây dựng và triển khai các tác nhân AI đối thoại. Bằng cách kết hợp LangGraph và Strands Agents, người dùng có thể tạo ra các tác nhân phân tích tài chính thông minh có thể cung cấp thông tin và khuyến nghị tài chính được cá nhân hóa. Những tác nhân này có thể phân tích dữ liệu tài chính, xác định xu hướng và mẫu, và cung cấp lời khuyên có thể hành động cho người dùng. Giải pháp này có tiềm năng cách mạng hóa cách dữ liệu tài chính được phân tích và sử dụng, cho phép các doanh nghiệp và cá nhân đưa ra quyết định tài chính thông minh hơn.

3

Các model không giới hạn trọng lượng tiếng Trung tốt nhất — và các đối thủ mạnh nhất của Mỹ

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

understandingai.org Đọc bài viết →

> Không có đủ nội dung được cung cấp để viết một bản tóm tắt. Tiêu đề "Các mô hình mở trọng lượng hàng đầu của Trung Quốc - và các đối thủ mạnh nhất của Mỹ" được đưa ra, nhưng nội dung thực tế lại thiếu. Vui lòng cung cấp nội dung để tôi có thể viết một bản tóm tắt rõ ràng và đầy đủ thông tin trong khoảng 150-200 từ.

4

Những người đam mê AI đang trong cuộc đua với thời gian, những người hoài nghi AI đang trong cuộc đua với entropy

AI enthusiasts are in a race against time, AI skeptics are in a race against entropy

Simon Willison Đọc bài viết →

Các nhà lãnh đạo trong ngành công nghệ đang đối mặt với một thách thức cấp bách khi công nghệ AI phát triển nhanh chóng. Một mặt, những người đam mê AI đang chấp nhận AI một cách tích cực và đạt được những đột phá đáng kể, dẫn đến những lợi thế cạnh tranh tiềm năng. Tuy nhiên, tốc độ phát triển nhanh chóng này cũng đặt ra rủi ro làm tổn hại đến độ tin cậy và khả năng bảo trì của hệ thống. Mặt khác, những người hoài nghi về AI ưu tiên sự thận trọng và ổn định lo lắng rằng việc vội vàng phát triển AI sẽ dẫn đến những vấn đề lâu dài và xói mòn niềm tin. Cả hai nhóm đều đúng, và vấn đề chính nằm ở sự thiếu sót của một vòng phản hồi tự nhiên kết nối họ. Khoảng cách này trong thực tế chia sẻ giữa những người đam mê và hoài nghi phải được giải quyết thông qua thiết kế tổ chức và lãnh đạo hiệu quả, đòi hỏi sự cân bằng giữa đổi mới và thận trọng.

5

Tự động hóa với tốc độ của Đầm lầy

Automation at the speed of Swamp

Changelog Đọc bài viết →

Trong một cuộc trò chuyện gần đây, Adam Jacob, người sáng lập System Initiative và tạo ra Swamp, đã thảo luận về tác động của các tác nhân AI đối với phát triển phần mềm. Với Swamp, một hệ thống tự động hóa các nhiệm vụ, đội ngũ 18 người của Jacob đã được giảm xuống còn năm người trong khi vẫn quản lý để xuất bản hệ thống 900 lần chỉ trong bốn tuần. Ông nhấn mạnh tầm quan trọng của kiến trúc phần mềm và thiết kế hướng lĩnh vực, cho rằng những yếu tố này hiện đang được ưu tiên hơn kỹ năng lập trình. Jacob cũng nhấn mạnh sự hồi sinh của kiểm thử Chấp nhận Người dùng (UAT), một phương pháp đã được sử dụng trước đây trong những năm 1990. Trong cuộc trò chuyện, một bản demo trực tiếp đã được trình diễn, nơi Swamp được sử dụng để tự động hóa các nhiệm vụ trên một hộp Proxmox, dẫn đến một kết quả đáng chú ý. Ngoài ra, Jacob cũng đề cập rằng ông không bao giờ chấp nhận các yêu cầu kéo (pull request) đối với Swamp, thể hiện cam kết của mình đối với sự phát triển của hệ thống.

6

Sau Orthogonality: Cơ quan đạo đức và sự phù hợp của AI

After Orthogonality: Virtue-Ethical Agency and AI Alignment

The Gradient Đọc bài viết →

Bài viết này khám phá khái niệm về việc căn chỉnh trí tuệ nhân tạo (AI) với quyền lực và giá trị của con người. Tác giả cho rằng những người hợp lý không có mục tiêu, mà thay vào đó căn chỉnh hành động của họ với các thực hành, là mạng lưới các hành động, khuynh hướng, tiêu chí đánh giá và tài nguyên. Để tạo ra các AI thực sự hỗ trợ quyền lực của con người, quá trình suy xét của các tác nhân AI phải chia sẻ một logic tương tự. Tác giả đề xuất rằng các khái niệm như "không gây hại" và "có thể sửa đổi" là tự nhiên hơn cho các tác nhân giải thích chúng như các động lực trong mạng lưới hành động, chứ không phải như mục tiêu hoặc quy tắc. Bài viết cũng giới thiệu khái niệm về "eudaimonia", hay sự thịnh vượng chủ động và hợp lý của con người, và cho rằng nó chỉ ra một cấu trúc suy xét khác với lý tính kết quả tiêu chuẩn. Tác giả đề xuất "lý tính eudaimonic" như một khuôn khổ hữu ích cho quyền lực và giá trị của AI được căn chỉnh với con người, nêu ra những lợi thế tiềm năng của nó về mặt ổn định và an toàn. Bài viết gợi ý rằng lý tính eudaimonic là một hình thức quyền lực tự nhiên có thể hiệu quả trong việc căn chỉnh AI với sự thịnh vượng của con người, và rằng cách tiếp cận này có thể giải quyết các vấn đề an toàn AI cổ điển và nghịch lý.

7

Cách Composer 2.5 của Cursor sử dụng tự chưng cất để vượt qua các LLM tiền phong trong mã hóa

How Cursor’s Composer 2.5 uses self-distillation to beat the frontier LLMs at coding

BD Tech Talks Đọc bài viết →

Cursor's Composer 2.5, một tác nhân mã hóa chuyên dụng, đã trở nên phổ biến средu các kỹ sư phần mềm nhờ tốc độ và hiệu quả về chi phí. Không giống như các mô hình tiền phong khổng lồ, đa mục đích, Composer 2.5 sử dụng một kỹ thuật gọi là tự chưng cất để xuất sắc trong các nhiệm vụ mã hóa. Phương pháp này liên quan đến việc đào tạo một mô hình nhỏ hơn để bắt chước đầu ra của một mô hình lớn hơn, cho phép nó học tập hiệu quả và hiệu suất. Composer 2.5 đạt được điều này thông qua một quá trình gọi là "RL mục tiêu với phản hồi văn bản", cung cấp các sửa lỗi cục bộ cho mô hình khi nó mắc lỗi. Kỹ thuật này giúp mô hình học hỏi từ sai lầm của nó mà không làm suy giảm khả năng tổng thể của nó. Bằng cách sử dụng tự chưng cất, Cursor đã tạo ra một mô hình có thể cạnh tranh với các mô hình lớn hơn như Opus 4.7 và GPT-5.5, nhưng với một phần nhỏ của chi phí. Mô hình định giá của Composer 2.5, với chi phí 0,50 đô la cho mỗi triệu token đầu vào và 2,50 đô la cho mỗi triệu token đầu ra, khiến nó trở

nên khả thi về mặt tài chính cho các đội nhỏ. Kết quả là, Composer 2.5 đã trở thành mặc định hàng ngày cho nhiều kỹ sư, cung cấp một giải pháp tiết kiệm hơn cho các nhiệm vụ mã hóa hàng ngày.

8 Nemotron 3.5 An toàn Nội dung: An toàn đa phương thức có thể tùy chỉnh cho Doanh nghiệp AI Toàn cầu

Nemotron 3.5 Content Safety: Customizable Multimodal Safety for Global Enterprise AI

Hugging Face Blog Đọc bài viết →

NVIDIA đã phát hành Nemotron 3.5 Content Safety, một bản cập nhật quan trọng cho ngăn xếp an toàn nội dung của họ. Phiên bản mới này kết hợp khả năng đa phương thức và đa ngôn ngữ, thực thi chính sách tùy chỉnh và lập luận có thể kiểm toán vào một model duy nhất. Nemotron 3.5 lấy một lời nhắc người dùng, một hình ảnh tùy chọn và một phản hồi trợ lý tùy chọn dưới dạng một cửa sổ ngữ cảnh duy nhất và tạo ra một phán quyết an toàn nhất quán trên đầu vào kết hợp. Model này duy trì phạm vi đào tạo rõ ràng 12 ngôn ngữ của các phiên bản trước và kế thừa sự khái quát hóa mạnh mẽ không có shot trên khoảng 140 ngôn ngữ từ model cơ sở Gemma 3. Phiên bản mới cũng bao gồm một thông số chính sách tùy chỉnh mà model lập luận khi tạo ra phán quyết của nó, thay vì hoàn toàn phụ thuộc vào phân loại tích hợp sẵn. Điều này cho phép có các chính sách an toàn được tùy chỉnh cho các ngành công nghiệp khác nhau, chẳng hạn như chăm sóc sức khỏe hoặc dịch vụ tài chính. Ngoài ra, Nemotron 3.5 có thể tạo ra các dấu vết lập luận có thể kiểm toán thông qua chế độ "THINK Mode" tùy chọn, cung cấp lập luận từng bước đằng sau phán quyết an toàn. NVIDIA cũng đang phát hành bộ dữ liệu an toàn của mình, một bộ dữ liệu đa phương thức và đa ngôn ngữ bao gồm các dấu vết lập luận an toàn được sử dụng để đào tạo model. Bộ dữ liệu này là một cột mốc quan trọng trong sự phát triển của các model an toàn nội dung, vì hầu hết các model an toàn mã nguồn mở không cung cấp tập dữ liệu đào tạo hoặc đánh giá. Nemotron 3.5 Content Safety được xây dựng trên model cơ sở Gemma 3 4B IT của Google, cung cấp khả năng lập luận mạnh mẽ về tầm nhìn và ngôn ngữ cũng như phạm vi đa ngôn ngữ rộng.

TIPS & TRICKS CHO DEV

Tự động hóa kiểm thử đơn vị

Vấn đề: Kiểm thử đơn vị tốn nhiều thời gian và nhân lực.

Cách làm: Sử dụng framework như Pytest, viết unit test tự động. Ví dụ: `pytest test_file.py`.

Đánh giá: Hiệu quả cao, giúp giảm thiểu thời gian kiểm thử, nên dùng cho dự án lớn.

Code Review Tự Động

Vấn đề: Kiểm tra code thủ công tốn thời gian và dễ xảy ra lỗi.

Cách làm: Sử dụng công cụ như SonarQube, thiết lập rule tự động. Ví dụ: `sonar-scanner -Dsonar.projectKey=project`.

Đánh giá: Hiệu quả cao, giúp phát hiện lỗi sớm, nên dùng cho dự án phức tạp.

Tự động hóa kiểm thử giao diện

Vấn đề: Kiểm thử giao diện người dùng tốn nhiều thời gian và khó khăn.

Cách làm: Sử dụng công cụ như Selenium, viết test tự động. Ví dụ: `driver.get("https://example.com")`.

Đánh giá: Hiệu quả cao, giúp giảm thiểu thời gian kiểm thử, nên dùng cho dự án có giao diện phức tạp.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tích hợp AI API vào ứng dụng**
2. Dev cần biết cách tích hợp AI API vào ứng dụng để tận dụng khả năng xử lý ngôn ngữ tự nhiên và tự động hóa các nhiệm vụ. Điều này giúp cải thiện hiệu suất và giảm thiểu thời gian phát triển. Tích hợp AI API cũng cho phép dev tạo ra các ứng dụng thông minh hơn và đáp ứng tốt hơn nhu cầu của người dùng.
3. Ví dụ, sử dụng API của LangGraph để tích hợp khả năng phân tích ngôn ngữ tự nhiên vào ứng dụng, giúp tự động hóa việc phân loại và trả lời các câu hỏi thường gặp.
4. Tip hoặc bước tiếp theo: Nghiên cứu các thư viện và framework hỗ trợ tích hợp AI API, như LangGraph, CrewAI, và bắt đầu với các dự án nhỏ để tích hợp AI vào ứng dụng của mình.