

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Người đẹp vì lụa, lúa tốt vì phân.”

— Tục ngữ Việt Nam

Môi trường, điều kiện và sự đầu tư đúng chỗ có thể giúp bất kỳ ai phát huy tối đa tiềm năng của mình.

TIN TỨC NỔI BẬT

1

Nhà máy Agent: Từ nguyên mẫu đến sản xuất - công cụ dành cho nhà phát triển và phát triển agent nhanh chóng

Agent Factory: From prototype to production—developer tools and rapid agent development

Microsoft Azure

[Đọc bài viết →](#)

Microsoft đã giới thiệu Agent Factory, một công cụ dành cho nhà phát triển được thiết kế để tối ưu hóa quá trình tạo và triển khai các tác nhân thông minh. Agent Factory cho phép nhà phát triển xây dựng, thử nghiệm và triển khai tác nhân một cách nhanh chóng, đẩy nhanh chu kỳ phát triển. Công cụ này được xây dựng trên nền tảng Microsoft Azure, tận dụng cơ sở hạ tầng đám mây và khả năng AI của nó. Agent Factory cung cấp một bộ công cụ và tính năng toàn diện giúp thúc đẩy sự phát triển của các tác nhân thông minh. Nhà phát triển có thể sử dụng các công cụ này để tạo, đào tạo và triển khai tác nhân, giúp dễ dàng tích hợp AI vào ứng dụng của họ. Nền tảng này hỗ trợ tạo mẫu và thử nghiệm nhanh, cho phép nhà phát triển nhanh chóng xác thực ý tưởng và lặp lại thiết kế của họ. Bằng cách sử dụng Agent Factory, nhà phát triển có thể tận dụng khả năng mở rộng và độ tin cậy của Microsoft Azure, đảm bảo rằng các tác nhân của họ được triển khai và quản lý một cách hiệu quả. Công cụ này được thiết kế để đơn giản hóa quá trình phát triển, cho phép nhà phát triển tập trung vào việc tạo ra các ứng dụng đổi mới tận dụng sức mạnh của AI.

2

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

Amazon Web Services (AWS) đã giới thiệu một phương pháp để xây dựng một tác nhân phân tích tài chính thông minh bằng cách sử dụng LangGraph và Strands Agents. Cách tiếp cận sáng tạo này cho phép người dùng tạo ra một công cụ phân tích tài chính phức tạp 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 người dùng trích xuất thông tin liên quan từ các báo cáo và tuyên bố tài chính. Mặt khác, Strands Agents là một framework để xây dựng các tác nhân thông minh có thể học hỏi và thích nghi với dữ liệu mới. Bằng cách kết hợp LangGraph và Strands Agents, người dùng có thể tạo ra một tác nhân phân tích tài chính có thể tự động trích xuất các chỉ số tài chính chính, xác định xu hướng và mẫu, và cung cấp thông tin chi tiết có thể hành động. Tác nhân thông minh này có thể được tích hợp với các hệ thống và công cụ tài chính khác nhau, khiến nó trở thành một tài sản quý giá cho các nhà phân tích tài chính, nhà đầu tư và tổ chức. Việc sử dụng LangGraph và Strands Agents trong phân tích tài chính mang lại một số lợi ích, bao gồm độ chính xác được cải thiện, hiệu quả tăng cao và khả năng ra quyết định được nâng cao.

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 →](#)

Bài viết thảo luận về các mô hình mở trọng lượng hàng đầu của Trung Quốc và những đối thủ mạnh nhất của Mỹ. Các mô hình mở trọng lượng đề cập đến robot hoặc máy móc hoạt động mà không có bất kỳ hạn chế trọng lượng nào, cho phép linh hoạt và đa năng hơn. Các mô hình Trung Quốc được nêu bật trong bài viết được coi là một trong những mô hình tốt nhất trong lớp của chúng, công nghệ tiên tiến và thiết kế sáng tạo. Những mô hình Trung Quốc này được so sánh với những đối thủ mạnh nhất của Mỹ, với bài viết cung cấp một so sánh về khả năng và tính năng của chúng. Các mô hình của Mỹ được biết đến với sự mạnh mẽ và tin cậy, khiến chúng trở thành một đối thủ đáng gờm đối với các đối thủ Trung Quốc. Bài viết nhằm cung cấp cho người đọc một hiểu biết toàn diện về trạng thái hiện tại của các mô hình mở trọng lượng, nêu bật điểm mạnh và điểm yếu của cả mô hình Trung Quốc và Mỹ. Bằng cách so sánh những mô hình này, người đọc có thể có cái nhìn sâu sắc về những tiến bộ mới nhất trong lĩnh vực robot và

trí tuệ nhân tạo (AI), và cập nhật những phát triển mới nhất trong lĩnh vực này, bao gồm cả các công nghệ như LLM, API, và framework, cũng như vai trò của developer trong việc tạo ra các mô hình này, và cách chúng xử lý token để đạt được hiệu suất cao.

4

Loa thông minh AI mới của OpenAI sẽ được bán với giá từ 300 đến 400 đô la

OpenAI's new AI smart speaker will reportedly sell for between \$300 and \$400

TechCrunch AI [Đọc bài viết →](#)

OpenAI được cho là đang phát triển một loa thông minh được hỗ trợ bởi AI, được mô tả là một thiết bị có hình dạng "donut" với vẻ ngoài "cao cấp" và được chế tạo từ "kim loại chất lượng cao". Thiết bị này, sẽ là hiện thân vật lý của ChatGPT, được thiết kế để di động và có thể được đặt ở nhiều vị trí khác nhau trong nhà. Các nguồn tin cho biết thiết bị sẽ có các "bộ phận di chuyển" riêng biệt, mặc dù mục đích của các tính năng này vẫn chưa rõ ràng. Loa thông minh này dự kiến sẽ có giá từ 300 đến 400 đô la, cao hơn đáng kể so với hầu hết các loa thông minh trong nhà trên thị trường. OpenAI đang hợp tác với LoveFrom, một studio thiết kế được thành lập bởi Jony Ive, để phát triển thiết bị này, dự kiến sẽ được phát hành vào năm 2027. Tuy nhiên, việc tham gia vào thị trường phần cứng của công ty có thể phải đối mặt với những thách thức, bao gồm cả vụ kiện từ Apple và khó khăn khi xâm nhập vào một thị trường có thể không mang lại lợi nhuận.

5

Hướng dẫn về lệnh cắt trong ứng dụng GitHub Copilot

A guide to slash commands in the GitHub Copilot app

GitHub Blog [Đọc bài viết →](#)

Người dùng ứng dụng GitHub Copilot có thể nâng cao trải nghiệm phát triển của mình với lệnh slash, một tính năng cung cấp lối tắt để quản lý phiên, điều hướng dự án và tùy chỉnh quy trình làm việc của họ. Những lối tắt văn bản này được nhập trực tiếp vào trình soạn thảo trò chuyện và bắt đầu bằng một dấu gạch chéo (/). Một menu tự động hoàn thành xuất hiện, hiển thị các lệnh có sẵn trong bối cảnh hiện tại. Lệnh slash trong ứng dụng GitHub Copilot mở rộng ý tưởng về các lối tắt nhanh, cung cấp một cách làm việc trực quan hơn với nền tảng phát triển được hỗ trợ bởi AI. Đối với người dùng quen thuộc với GitHub Copilot CLI, sự khác biệt chính nằm ở quy trình làm việc, vì

lệnh slash CLI được thiết kế xung quanh cách tiếp cận đầu cuối. Ngược lại, lệnh slash của ứng dụng tập trung vào việc tối ưu hóa các tương tác trong trình soạn thảo trò chuyện. Bằng cách tận dụng những lỗi tắt này, các developer có thể tăng năng suất và hiệu quả trong quy trình làm việc của họ.

6

Từ các dự án nguồn mở nổi tiếng đến OpenAI

From open source hits to OpenAI

Changelog

[Đọc bài viết →](#)

Trong tập này, Max Stoiber, một developer làm việc trên thư mục plugin và nền tảng ứng dụng của ChatGPT tại OpenAI, chia sẻ những nhận xét của mình về ngành công nghệ. Ông thảo luận về các dự án mã nguồn mở thường bị bỏ qua, chẳng hạn như react-boilerplate và styled-components, những dự án đã có những đóng góp đáng kể cho cảnh quan công nghệ. Stoiber cũng đề cập đến việc GitHub mua lại Spectrum, điều đã đóng vai trò quan trọng trong việc định hình GitHub Discussions. Ngoài ra, ông cũng nói về sự phát triển của Stellate, một bộ nhớ đệm GraphQL đã được cả Shopify và The Guild mua lại. Cuộc trò chuyện cũng khám phá khái niệm về các ứng dụng ChatGPT như một bề mặt mới cho phát triển phần mềm.

7

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, quan niệm này dựa trên khả năng của các mô hình có thể mở rộng hiệu quả trên phần cứng hiện có, chứ không phải là một cách tiếp cận có suy nghĩ về trí tuệ. Cách tiếp cận đa phương thức, kết hợp nhiều phương thức để đạt được trí tuệ tổng quát, không thể thành công trong thời gian tới. Thay vào đó, các nhà nghiên cứu nên tập trung vào việc phát triển AGI coi sự thể hiện và tương tác với môi trường là yếu tố chính, với việc xử lý tập trung vào phương thức xuất hiện như một kết quả. 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 bắt nguồn từ thế giới vật lý, chẳng hạn như sửa chữa một chiếc xe hơi hoặc tháo một nút thắt. 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 heuristic để dự đoán token, chứ không phải là một sự hiểu biết thực sự về thực tại.

8

OpenAI cho rằng vụ kiện bí mật thương mại của Apple là 'thối rữa từ lõi'

OpenAI says Apple's trade secrets lawsuit is 'rotten to its core'

The Verge AI

[Đọc bài viết →](#)

OpenAI đã nộp đơn yêu cầu bác bỏ vụ kiện của Apple, cáo buộc gã khổng lồ công nghệ này đã mô tả sai hành động của các cựu nhân viên Apple đã gia nhập OpenAI. Apple đã cáo buộc những nhân viên này đã đánh cắp bí mật thương mại, nhưng OpenAI cho rằng thông tin đang được tranh cãi là thông tin chung và không phải là bí mật. Công ty khởi nghiệp AI này lập luận rằng Apple đã không duy trì bí mật và rằng vụ kiện là "không có căn cứ". OpenAI cũng cáo buộc Apple là "cẩn thận, hung hăng và kỳ lạ là cá nhân" trong việc theo đuổi vụ việc này. Công ty cho rằng họ đang xây dựng điều gì đó mới và khác biệt so với sản phẩm của Apple, và họ có quyền thuê những kỹ sư và developer giỏi nhất. OpenAI đang tìm cách bác bỏ vụ kiện, với một phiên tòa dự kiến vào ngày 1 tháng 10. Trong khi đó, Apple đã yêu cầu một lệnh cấm sơ bộ để ngăn chặn OpenAI và các nhân viên tiếp cận hoặc tiết lộ thông tin bí mật cáo buộc.

TIPS & TRICKS CHO DEV

Context Window Management

Vấn đề: Quản lý context window hiệu quả để tránh mất thông tin quan trọng.

Cách làm: Sử dụng kỹ thuật **context window management** để điều chỉnh kích thước cửa sổ context. Ví dụ, sử dụng lệnh `--max-context-length` trong CLI để thiết lập kích thước tối đa.

Đánh giá: Hiệu quả trong việc quản lý thông tin, đặc biệt khi làm việc với dữ liệu dài.

Long-Context Modeling

Vấn đề: Xử lý các mô hình có context dài và phức tạp.

Cách làm: Áp dụng kỹ thuật **long-context modeling** để mô hình hóa các mối quan hệ dài. Ví dụ, sử dụng mô hình `Transformer` với kỹ thuật **attention**

mechanism.

Đánh giá: Hiệu quả trong việc xử lý các mối quan hệ dài và phức tạp.

Memory Optimization

Vấn đề: Tối ưu hóa bộ nhớ để giảm thiểu tiêu thụ tài nguyên.

Cách làm: Sử dụng kỹ thuật **memory optimization** để giảm thiểu sử dụng bộ nhớ. Ví dụ, sử dụng lệnh `--memory-optimization` trong CLI để bật tính năng tối ưu hóa bộ nhớ.

Đánh giá: Hiệu quả trong việc giảm thiểu tiêu thụ tài nguyên và cải thiện hiệu suất.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tích hợp AI API vào ứng dụng

2. Tích hợp AI API vào ứng dụng là một chủ đề quan trọng vì nó giúp các nhà phát triển ứng dụng tận dụng khả năng của trí tuệ nhân tạo để cải thiện chức năng và trải nghiệm người dùng. Điều này cho phép các ứng dụng trở nên thông minh hơn và có khả năng xử lý các tác vụ phức tạp. Việc tích hợp AI API cũng giúp giảm thiểu thời gian và công sức phát triển.

3. Ví dụ, một ứng dụng có thể sử dụng API của Google Cloud Vision để phân tích hình ảnh và nhận diện đối tượng, hoặc sử dụng API của Dialogflow để tạo ra chatbot thông minh.

4. Tip hoặc bước tiếp theo: Để tích hợp AI API vào ứng dụng, bạn nên bắt đầu bằng việc xác định nhu cầu và mục tiêu của ứng dụng, sau đó lựa chọn AI API phù hợp và thực hiện tích hợp bằng cách sử dụng các thư viện và framework hỗ trợ như TensorFlow hoặc PyTorch.

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

Nguồn: Google News · Groq AI