

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Tiên học lễ, hậu học văn.”

— Tục ngữ Việt Nam

Đạo đức và nhân cách phải được xây dựng trước kiến thức — người có học vấn mà thiếu lễ nghĩa thì cũng vô ích.

TIN TỨC NỔI BẬT

Giới thiệu Open Agent Specification (Agent Spec): Một cách biểu diễn thống nhất cho AI Agent

1

Introducing the Open Agent Specification (Agent Spec): A Unified Representation for AI Agents

Oracle Blogs

[Đọc bài viết →](#)

Oracle đã giới thiệu Open Agent Specification (Agent Spec), một biểu diễn thống nhất cho các tác nhân AI. Agent Spec nhằm cung cấp một khuôn khổ tiêu chuẩn hóa để mô tả và tương tác với các tác nhân AI, cho phép tích hợp và giao tiếp liền mạch trên các hệ thống và nền tảng khác nhau. Thông số kỹ thuật này được thiết kế để mở và có thể mở rộng, cho phép các nhà phát triển tạo và chia sẻ các tác nhân AI của riêng họ và tích hợp chúng với các ứng dụng khác nhau. Agent Spec được xây dựng trên các công nghệ và tiêu chuẩn hiện có, chẳng hạn như JSON và HTTP, để đảm bảo khả năng tương tác và dễ dàng áp dụng. Nó cung cấp một cách rõ ràng và concis để biểu diễn các tác nhân AI, bao gồm khả năng, ý định và hành vi của chúng. Biểu diễn thống nhất này cho phép các nhà phát triển tập trung vào việc xây dựng và tích hợp các tác nhân AI, thay vì lo lắng về các vấn đề tương thích. Bằng cách áp dụng Agent Spec, các nhà phát triển có thể tạo ra các hệ thống AI phức tạp và kết nối hơn, và các tổ chức có thể tận dụng một loạt các tác nhân AI để thúc đẩy đổi mới và tăng trưởng. Agent Spec là một dự án mã nguồn mở, và Oracle mời cộng đồng đóng góp và tham gia vào quá trình phát triển của nó.

2

Tăng tốc phát triển với Amazon Bedrock AgentCore MCP server | Trí tuệ nhân tạo

Accelerate development with the Amazon Bedrock AgentCore MCP server | Artificial Intelligence

Amazon Web Services (AWS)

[Đọc bài viết →](#)

Amazon Web Services (AWS) đã giới thiệu máy chủ Amazon Bedrock AgentCore MCP, được thiết kế để tăng tốc phát triển trong lĩnh vực trí tuệ nhân tạo (AI). Máy chủ AgentCore MCP là một phần của nền tảng Amazon Bedrock, nhằm mục đích đơn giản hóa việc phát triển và triển khai các model AI. Máy chủ AgentCore MCP là một máy chủ hiệu suất cao cho phép các nhà phát triển xây dựng, đào tạo và triển khai các model AI một cách hiệu quả hơn. Nó cung cấp một môi trường có thể mở rộng và bảo mật cho phát triển AI, cho phép các nhà phát triển tập trung vào xây dựng các ứng dụng AI sáng tạo. Với máy chủ AgentCore MCP, các nhà phát triển có thể tận dụng chuyên môn của Amazon trong lĩnh vực AI và học máy để tăng tốc quá trình phát triển của họ. Máy chủ này được tích hợp với các dịch vụ AWS khác, giúp dễ dàng triển khai và quản lý các model AI trên đám mây. Máy chủ Amazon Bedrock AgentCore MCP là một công cụ quý giá cho các nhà phát triển muốn xây dựng và triển khai các model AI một cách nhanh chóng và hiệu quả. Bằng cách cung cấp một môi trường có hiệu suất cao và có thể mở rộng, nó cho phép các nhà phát triển tập trung vào đổi mới và đưa các dự án AI của họ vào cuộc sống.

GPT-5 của OpenAI đã ra mắt: Phân tích chuyên sâu System Card cho AI thông minh hơn, an toàn hơn và nhanh hơn

3

OpenAI's GPT-5 Is Here: A Deep Dive Into the System Card for AI That's Smarter, Safer, and Faster

Medium

[Đọc bài viết →](#)

OpenAI đã công bố việc phát hành GPT-5, một bản cập nhật quan trọng cho hệ thống AI của họ. Mô hình mới này tự hào với các khả năng được cải tiến, bao gồm trí tuệ tăng cường, các tính năng an toàn được nâng cao và tốc độ xử lý được tăng tốc. Một khía cạnh quan trọng của GPT-5 là khả năng học hỏi và thích nghi hiệu quả hơn, cho phép nó giải quyết các nhiệm vụ phức tạp một cách dễ dàng hơn. Thẻ hệ thống cho GPT-5 được thiết kế với sự an toàn trong tâm trí, kết hợp các biện pháp khác nhau để ngăn chặn việc lạm dụng tiềm năng. Các tính năng này nhằm giảm thiểu rủi ro liên quan đến các hệ thống AI tiên tiến, đảm bảo một môi trường an toàn và được kiểm soát hơn. Hiệu suất được cải tiến của GPT-5 được cho là nhờ vào kiến trúc tiên

tiến và phương pháp đào tạo của nó. Khả năng xử lý thông tin nhanh hơn và chính xác hơn của mô hình cho phép nó xử lý một loạt các nhiệm vụ rộng lớn hơn, từ tạo văn bản đến cung cấp các phản hồi sâu sắc. Kết quả là, GPT-5 có khả năng tạo ra tác động đáng kể đối với các ngành công nghiệp khác nhau, bao gồm chăm sóc sức khỏe, tài chính và giáo dục, nơi các giải pháp được hỗ trợ bởi AI ngày càng được yêu cầu.

4

Profiling trong PyTorch (Phần 3): Attention là tất cả những gì bạn profile

Profiling in PyTorch (Part 3): Attention is all you profile

Hugging Face Blog

[Đọc bài viết →](#)

Trong phần thứ ba của "Profiling in PyTorch", một loạt bài tập trung vào việc đọc các dấu vết của profiler để thúc đẩy tối ưu hóa, tác giả khám phá cơ chế chú ý, một thành phần cơ bản của kiến trúc Transformer. Mặc dù có độ phức tạp thời gian bậc hai nổi tiếng, chú ý có thể được tối ưu hóa bằng cách sử dụng các kỹ thuật khác nhau. Tác giả phân tích bốn triển khai khác nhau của chú ý: chú ý đơn giản, chú ý với mặt nạ nguyên nhân tại chỗ, chú ý dot sản phẩm có tỷ lệ và chú ý với kernel hiệu quả. Bằng cách phân tích các dấu vết của profiler, tác giả xác định các cơ hội để tối ưu hóa, chẳng hạn như thay thế các hoạt động ngoài chỗ bằng các hoạt động tại chỗ, điều này có thể giảm đáng kể số lượng hoạt động CPU và loại bỏ các bản sao bộ nhớ không cần thiết trên GPU. Những phát hiện này có ý nghĩa đối với các model lớn, nơi các hoạt động chú ý được lặp lại nhiều lần, và có thể dẫn đến lợi ích hiệu suất đáng kể.

5

Phát hiện trong một repo không phải là một security posture hoàn chỉnh

Detection in one repo isn't a security posture

Sourcegraph Blog

[Đọc bài viết →](#)

Một phát hiện gần đây nhấn mạnh tầm quan trọng của các biện pháp bảo mật toàn diện vượt ra ngoài việc phát hiện riêng lẻ từng kho lưu trữ. Trong khi việc phát hiện ra các điểm yếu trong một kho lưu trữ riêng lẻ là một bước quan trọng, thì nó chỉ là một phần của một tư thế bảo mật rộng lớn hơn. Một tư thế bảo mật bao gồm không chỉ việc xác định các vấn đề mà còn hiểu sự hiện diện của chúng trên tất cả các

kho lưu trữ và thực hiện các biện pháp chủ động để giải quyết chúng trước khi chúng lan rộng. Cách tiếp cận này nhấn mạnh nhu cầu vượt ra ngoài việc phát hiện bị cô lập và thay vào đó tập trung vào việc tạo ra một khuôn khổ bảo mật mạnh mẽ có thể xác định và giảm thiểu các mối đe dọa tiềm năng trên toàn hệ thống. Bằng cách làm như vậy, các tổ chức có thể giảm thiểu rủi ro của các điểm yếu lan truyền và gây ra thiệt hại rộng lớn. Cuối cùng, một tư thế bảo mật không chỉ là về việc phát hiện các vấn đề trong một kho lưu trữ, mà còn về việc duy trì một tư thế cảnh giác và chủ động đối với bảo mật trên tất cả các khía cạnh của hệ thống.

6

Meta tắt tính năng Instagram cho phép người dùng tạo AI deepfake từ các tài khoản công khai

Meta turns off the Instagram feature that let users make AI deepfakes of public accounts

The Verge AI

[Đọc bài viết →](#)

Meta đã thông báo rằng họ sẽ tắt một tính năng trên Instagram cho phép người dùng tạo ra hình ảnh được tạo bởi AI dựa trên nội dung từ các tài khoản công khai bằng cách gắn thẻ chúng. Tính năng này, được giới thiệu vào đầu tuần, đã nhận được sự phản ứng dữ dội từ người dùng và các tổ chức, bao gồm Trung tâm Quốc gia về Khai thác Tình dục và Hội Diễn viên Màn ảnh. Các nhà phê bình cho rằng tính năng này đã xói mòn quyền của người dùng đối với hình ảnh của chính họ và tạo ra rủi ro cho các trò lừa đảo và khai thác. Meta cho biết ý định của họ là cung cấp một công cụ sáng tạo và trao cho người dùng quyền kiểm soát nội dung của họ, nhưng thừa nhận rằng tính năng này "đã bỏ lỡ mục tiêu." Người dùng muốn chọn không tham gia tính năng này phải đào sâu vào cài đặt, điều này được coi là không thể chấp nhận được bởi một số người. Kết quả là, tính năng này đã bị tắt hoàn toàn.

7

AI Agent tạo investment memo chỉ trong 90 giây, tiết kiệm hàng giờ làm việc cho analyst

LangChain Blog

[Đọc bài viết →](#)

Một đội đã phát triển một tác nhân trí tuệ nhân tạo (AI) có thể tạo ra một phiên bản đầu tiên của bản ghi nhớ đầu tư chỉ trong 90 giây, giúp các nhà phân tích tiết kiệm hàng giờ thời gian nghiên cứu và viết. Tác nhân này sử dụng sự kết hợp của các API và công cụ để thu thập

thông tin từ các nguồn chính, bao gồm Crunchbase, PitchBook và trang IR. Nó tạo ra một bản ghi nhớ có căn cứ với bảy phần: Tóm tắt, Đội ngũ, Tài chính, Sản phẩm, Thị trường, Rủi ro và Luận điểm, bao gồm một khuyến nghị một dòng. Tác nhân này hoạt động ở hai giai đoạn: các nút nghiên cứu thu thập bằng chứng song song và một nút tổng hợp lắp ráp bản ghi nhớ từ các đầu ra nghiên cứu. Tác nhân này được cung cấp năng lượng bởi Perplexity Agent API và các công cụ web_search và finance_search tích hợp, với LangGraph để điều phối và LangSmith để đánh giá. Đầu ra của tác nhân này có thể được sử dụng làm điểm khởi đầu cho các nhà phân tích để hoàn thiện và tinh chỉnh, thay vì bắt đầu từ đầu. Đội cũng đã phát triển một hệ thống đánh giá để đánh giá hiệu suất của các nhà cung cấp tìm kiếm khác nhau, có thể được tùy chỉnh để phù hợp với các trường hợp sử dụng cụ thể.

8

Cách GitHub gán quyền owner bền vững cho mọi repository

How GitHub gave every repository a durable owner

GitHub Blog

[Đọc bài viết →](#)

GitHub, một nền tảng hàng đầu cho các developer, đã triển khai một giải pháp để giải quyết vấn đề lâu dài về quyền sở hữu repository. Công ty này có hơn 14.000 repository, nhưng ít hơn một nửa trong số đó có quyền sở hữu rõ ràng. Sự thiếu hụt này dẫn đến khó khăn trong việc quản lý và bảo mật repository, đặc biệt là trong nỗ lực khắc phục vấn đề quét bí mật. Để giải quyết vấn đề này, các đội kỹ sư và bảo mật của GitHub đã làm việc cùng nhau để xác thực quyền sở hữu cho mọi repository hoạt động trong thời gian 45 ngày. Kết quả là khoảng 8.000 repository không hoạt động đã được lưu trữ và quyền sở hữu đã được thiết lập cho các repository hoạt động còn lại. Nền tảng mới này về quyền sở hữu đã cho phép GitHub cải thiện trải nghiệm của developer và các biện pháp bảo mật, đặc biệt là trong các lĩnh vực như DevSecOps. Sự thành công của công ty trong việc giải quyết vấn đề này cũng đã được công nhận bởi Gartner, khi họ xếp GitHub vào vị trí Leader lần thứ hai liên tiếp.

TIPS & TRICKS CHO DEV

Tối ưu hóa multi-agent

Vấn đề: Xử lý task phức tạp đòi hỏi nhiều AI agents.

Cách làm: Sử dụng LangGraph, CrewAI, AutoGen, phối hợp nhiều AI agents. Ví

dụ: `langgraph --task complex_task --agents 3`.

Đánh giá: Hiệu quả cho task lớn, nhưng tăng độ phức tạp.

Cài đặt phiData

Vấn đề: Cần quản lý dữ liệu đầu vào cho nhiều AI agents.

Cách làm: Sử dụng phiData, cấu hình dữ liệu đầu vào cho từng agent. Ví dụ:

```
phidata --input data.csv --agent agent1
```

Đánh giá: Giúp quản lý dữ liệu hiệu quả, giảm sai sót.

Tích hợp AutoGen

Vấn đề: Cần tự động hóa việc tạo code cho các AI agents.

Cách làm: Sử dụng AutoGen, tích hợp vào pipeline phát triển. Ví dụ: `autogen --template template.py --output output.py`.

Đánh giá: Tự động hóa hiệu quả, tiết kiệm thời gian phát triển.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tích hợp AI API vào ứng dụng

Dev cần biết cách tích hợp AI API vào ứng dụng để tận dụng khả năng xử lý tự nhiên ngôn ngữ, hình ảnh và dữ liệu khác. Điều này giúp tăng cường tính tương tác và thông minh cho ứng dụng.

2. Việc tích hợp AI API giúp dev tạo ra các ứng dụng thông minh hơn, tự động hóa các tác vụ và cải thiện trải nghiệm người dùng.

3. Ví dụ, sử dụng AI API để phân tích hình ảnh và nhận diện đối tượng, sau đó đưa ra các gợi ý cho người dùng.

4. Tip hoặc bước tiếp theo: Sử dụng các nền tảng như LangGraph hoặc CrewAI để tích hợp AI vào ứng dụng và khám phá các khả năng của AI API trong việc cải thiện ứng dụng của bạn.

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

Nguồn: Google News · Groq AI