

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Miệng nam mô lòng dao kiếm.”

— Tục ngữ Việt Nam

Đừng nhìn bề ngoài mà phán xét người — hãy quan sát hành động và đánh giá người qua việc làm thực tế.

TIN TỨC NỔI BẬT

1

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

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 cách biểu diễn thống nhất cho các AI agent. Agent Spec nhằm mục đích cung cấp một framework tiêu chuẩn để mô tả và tương tác với các AI agent, cho phép giao tiếp và tích hợp liền mạch trên các hệ thống và nền tảng khác nhau. Agent Spec được thiết kế để không phụ thuộc vào ngôn ngữ và nền tảng, cho phép các developer tạo ra các AI agent có thể dễ dàng tích hợp với nhiều hệ thống và ứng dụng khác nhau. Specification này cung cấp một bộ từ vựng và cấu trúc chung để mô tả các khả năng, mục tiêu và hành vi của AI agent, giúp việc phát triển và triển khai các giải pháp ứng dụng AI trở nên dễ dàng hơn. Bằng cách áp dụng Agent Spec, các developer có thể tạo ra các hệ thống AI có khả năng tương tác và mở rộng tốt hơn, giảm thiểu sự phức tạp và phân mảnh liên quan đến quá trình phát triển AI hiện tại. Agent Spec là mã nguồn mở và được cộng đồng phát triển, đảm bảo rằng nó vẫn linh hoạt và có thể thích ứng với các nhu cầu phát triển của ngành AI. Specification này có tiềm năng đẩy nhanh quá trình phát triển và triển khai các giải pháp ứng dụng AI trên nhiều ngành và lĩnh vực khác nhau.

2

Agent Factory: Kết nối các agent, app và data với các open standard 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 được thiết kế để kết nối các agent, app và data bằng cách sử dụng các open standard. Nền tảng này tận dụng hai standard chính: MCP (Microsoft Cloud Protocol) và A2A (Application to Application). MCP là một protocol dựa trên cloud cho phép giao tiếp an toàn và hiệu quả giữa các ứng dụng, trong khi A2A tạo điều kiện trao đổi data liền mạch giữa các hệ thống khác nhau. Agent Factory nhằm mục đích đơn giản hóa quy trình tích hợp bằng cách cung cấp một cách thức tiêu chuẩn để kết nối nhiều agent, app và nguồn data khác nhau. Điều này cho phép các developer tập trung vào việc xây dựng ứng dụng thay vì dành thời gian cho các tích hợp phức tạp. Các open standard của nền tảng cho phép linh hoạt và khả năng tương tác cao hơn, giúp việc tích hợp với các hệ thống và ứng dụng hiện có trở nên dễ dàng hơn. Bằng cách sử dụng Agent Factory, các developer có thể tạo ra các ứng dụng hiệu quả và có khả năng mở rộng tốt hơn, dễ dàng thích ứng với các nhu cầu kinh doanh thay đổi. Các open standard của nền tảng cũng thúc đẩy sự hợp tác và đổi mới, vì các developer có thể xây dựng dựa trên công việc hiện có và chia sẻ kiến thức với cộng đồng.

3

Tăng tốc phát triển với Amazon Bedrock AgentCore MCP server | Artificial Intelligence

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 Amazon Bedrock AgentCore MCP server, được thiết kế để tăng tốc phát triển trong lĩnh vực Artificial Intelligence (AI). AgentCore MCP server là một phần của nền tảng Bedrock, cho phép các developer xây dựng, train và deploy các AI model ở quy mô lớn. AgentCore MCP server là một server hiệu năng cao, cung cấp một môi trường an toàn và có khả năng mở rộng cho việc phát triển AI. Nó được tối ưu hóa cho các workload machine learning (ML) và hỗ trợ một loạt các AI framework và tool. Với AgentCore MCP server, các developer có thể nhanh chóng và dễ dàng deploy các AI model, giảm thời gian và công sức cần thiết để phát triển và deploy các ứng dụng AI. Amazon Bedrock AgentCore MCP

server được tích hợp với các AWS service khác, cho phép các developer tận dụng toàn bộ khả năng của AWS để xây dựng và deploy các ứng dụng AI. Điều này bao gồm quyền truy cập vào mạng lưới rộng lớn các compute resource, storage và database của AWS. Bằng cách sử dụng AgentCore MCP server, các developer có thể tăng tốc quá trình phát triển và deploy AI của mình, giúp họ đưa các ứng dụng ứng dụng AI ra thị trường nhanh hơn.

4

Harness là tất cả những gì bạn cần (chủ yếu)

The harness is all you need (mostly)

GitHub Blog

[Đọc bài viết →](#)

GitHub đang cung cấp một loạt các tài nguyên và insight về artificial intelligence và machine learning trên toàn bộ hệ sinh thái của mình và trong ngành công nghiệp rộng lớn hơn. Nền tảng này đang khám phá các khả năng và lợi ích của việc AI code generation, có thể cải thiện developer experience. GitHub cũng chia sẻ các best practice để xây dựng

5

Từ những thành công open source đến OpenAI

From open source hits to OpenAI

Changelog

[Đọc bài viết →](#)

Trong một cuộc trò chuyện gần đây, Max Stoiber, một developer làm việc trên thư viện plugin và nền tảng ứng dụng của ChatGPT tại OpenAI, đã thảo luận về các dự án mã nguồn mở và tác động của chúng đối với ngành công nghệ. Ông đã đề cập đến các dự án đáng chú ý như react-boilerplate và styled-components, những dự án đã có ảnh hưởng đáng kể. Stoiber cũng đề cập đến việc GitHub mua lại Spectrum, điều này đã dẫn đến sự phát triển của GitHub Discussions. Ngoài ra, ông cũng nói về sự phát triển của Stellate, một bộ nhớ đệm GraphQL đã được Shopify và The Guild mua lại. Stoiber so sánh các ứng dụng ChatGPT với một bề mặt mới cho phần mềm, gợi ý rằng chúng cung cấp một cách mới để tương tác với công nghệ. Cuộc trò chuyện cũng nhấn mạnh các nhà tài trợ khác nhau, bao gồm Coder.com, WorkOS, Notion, Fly.io và Changelog.com, những nhà tài trợ cung cấp môi trường an toàn, công cụ xác thực, tác nhân tùy chỉnh và dịch vụ triển khai.

6

Sau Orthogonality: Agency theo đạo đức đức hạnh và AI Alignment

After Orthogonality: Virtue-Ethical Agency and AI Alignment

The Gradient

[Đọc bài viết →](#)

Trong bài viết này, tác giả cho rằng những người có lý trí không có mục tiêu, mà thay vào đó là sắp xếp hành động của họ theo các thực hành, những mạng lưới gồm các hành động, khuynh hướng và tiêu chí đánh giá. Tác giả đề xuất rằng các tác nhân trí tuệ nhân tạo (AI) nên được thiết kế để chia sẻ logic dựa trên thực hành này nhằm hỗ trợ và cộng tác một cách chân thành với khả năng hành động của con người. Cách tiếp cận này rất quan trọng để đảm bảo AI phù hợp với các thuộc tính an toàn như minh bạch, hữu ích và vô hại. Tác giả giới thiệu khái niệm "tư duy lý trí eudaimonic", được lấy cảm hứng từ ý tưởng về sự thịnh vượng của con người (eudaimonia). Họ lập luận rằng hình thức hoạt động lý trí này là một khuôn khổ hữu ích cho khả năng hành động và giá trị của AI được đồng bộ hóa với con người, vì nó không phân biệt giữa phương tiện và mục đích, và ổn định, an toàn hơn so với các phương pháp truyền thống theo hệ quả hoặc nghĩa vụ. Tác giả tuyên bố rằng tư duy lý trí eudaimonic là một hình thức khả năng hành động tự nhiên có thể hiệu quả ngay cả theo tiêu chuẩn hệ quả, và nó có thể là chìa khóa để giải quyết các vấn đề và nghịch lý an toàn AI cổ điển.

7

Tại sao Trung Quốc lại "cho không" các AI model tốt nhất của mình

Why China is giving away its best AI models

The Verge AI

[Đọc bài viết →](#)

Công ty Moonshot AI của Trung Quốc đã phát hành một mô hình AI mạnh mẽ có tên là Kimi K3, được cho là có thể vượt trội hơn một số hệ thống tốt nhất được xây dựng bởi các công ty Mỹ với chi phí thấp hơn. Trọng số của mô hình này đang được cung cấp miễn phí, làm dấy lên lo ngại về sự thống trị của các mô hình AI đóng của Mỹ. Các mô hình trọng số mở, như Kimi K3, mang lại cho các nhà phát triển quyền kiểm soát và linh hoạt lớn hơn, cho phép họ tùy chỉnh và xây dựng các sản phẩm mới mà không phụ thuộc vào một nhà cung cấp duy nhất. Cách tiếp cận này thường rẻ hơn và dễ tiếp cận hơn so với các hệ thống độc quyền. Mặc dù các mô hình trọng số mở không hoàn toàn "mở" theo nghĩa truyền thống, nhưng chúng vẫn có thể mang lại nhiều lợi ích và cơ hội đáng kể cho các công ty. Bằng cách phát hành trọng số mô

hình, các công ty có thể khuyến khích nhiều nhà phát triển sử dụng mô hình của họ, dẫn đến một hệ sinh thái các công cụ và cơ sở hạ tầng được xây dựng xung quanh chúng. Điều này có thể giúp một mô hình trở thành "tiêu chuẩn" thực" và tạo ra một lợi thế cạnh tranh. Việc phát hành Kimi K3 đã làm dấy lên lo ngại cho các gã khổng lồ AI của Mỹ, như OpenAI, Google và Anthropic, vì nó có thể làm thay đổi trọng tâm của ngành công nghiệp khỏi các nền tảng độc quyền. Tác động của các mô hình trọng số mở đối với thị trường vẫn còn phải được xem xét, nhưng chúng đang buộc các công ty Mỹ phải suy nghĩ lại cách tiếp cận của họ đối với phát triển và phân phối AI.

8

Tại sao các hệ thống nhận thức ứng dụng AI đang định nghĩa lại Radar và Electronic Warfare

Why AI-Driven Cognitive Systems Are Redefining Radar and Electronic Warfare

IEEE Spectrum

[Đọc bài viết →](#)

Các hệ thống radar truyền thống và chiến tranh điện tử đang phải đối mặt với một thách thức đáng kể do các mối đe dọa mới xuất hiện từ các bộ phát mode-agile hoạt động ở các chế độ không truyền thống. Các hệ thống này dựa trên thư viện mối đe dọa tĩnh, không thể phù hợp với các mối đe dọa không xác định. Để giải quyết vấn đề này, các hệ thống RF nhận thức đang được phát triển, sử dụng các kỹ thuật trí tuệ nhân tạo (AI) và học máy (ML). Các hệ thống này cho phép nhận thức, lý luận và phản hồi tự động với các mối đe dọa không xác định trong phổ RF. Một bài báo trắng xem xét kiến trúc của các hệ thống radar và chiến tranh điện tử nhận thức AI/ML, bao gồm các khối chức năng chính như thu thập RF, phân tích AI và tổng hợp sóng. Bài báo cũng thảo luận về các thách thức trong việc đào tạo các hệ thống này, bao gồm việc thu thập tập dữ liệu tín hiệu và thực hiện kiểm tra, và mô tả cách các thử nghiệm vòng kín có thể phát triển và cải thiện các thuật toán AI/ML để chống lại các mối đe dọa không xác định. Sự phát triển này nhằm nâng cao hiệu quả của các hệ thống radar và chiến tranh điện tử khi đối mặt với các mối đe dọa mới xuất hiện.

TIPS & TRICKS CHO DEV

Cài đặt Context Window

Vấn đề: Quản lý context window để tối ưu hóa hiệu suất của mô hình AI.

Cách làm: Sử dụng kỹ thuật chunking để chia đầu vào thành các đoạn nhỏ, ví dụ

như trong câu lệnh `transformers BartForConditionalGeneration`.

Đánh giá: Hiệu quả khi xử lý văn bản dài, giúp giảm thiểu việc mất thông tin.

Tối ưu Long-Context

Vấn đề: Xử lý thông tin dài trong các ứng dụng AI.

Cách làm: Áp dụng kỹ thuật attention mechanism, ví dụ như trong lệnh

`huggingface/transformers`.

Đánh giá: Tăng cường khả năng giữ thông tin, phù hợp cho các nhiệm vụ yêu cầu thông tin dài.

Quản lý Memory

Vấn đề: Quản lý bộ nhớ hiệu quả cho các ứng dụng AI.

Cách làm: Sử dụng kỹ thuật model pruning để giảm kích thước mô hình, ví dụ như trong lệnh `torch.nn.utils.prune`.

Đánh giá: Giúp giảm thiểu sử dụng bộ nhớ, tăng tốc độ xử lý cho các ứng dụng AI.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Để phát triển ứng dụng AI hiệu quả, các dev cần biết cách tối ưu chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM). Việc này giúp giảm thiểu chi phí vận hành và cải thiện trải nghiệm người dùng. Điều này đặc biệt quan trọng khi triển khai ứng dụng trên quy mô lớn.

3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và tăng tốc độ xử lý. Code snippet minh họa:

```
model = LLM.from_pretrained('base_model'); model.fine_tune(...)
```

4. Tip hoặc bước tiếp theo: Để tối ưu hóa hiệu năng, hãy xem xét việc sử dụng các mô hình AI được đào tạo trước và điều chỉnh chúng cho use case cụ thể của bạn. Điều này có thể giúp giảm thiểu thời gian đào tạo và cải thiện hiệu suất của ứng dụng.

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

Nguồn: Google News · Groq AI