BIS - MT · 04/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Không thầy đố mày làm nên."

— Tục ngữ Việt Nam

Người thầy, người mentor đóng vai trò quan trọng trong sự phát triển — biết tìm kiếm và trân trọng sự chỉ dẫn là khôn ngoan.

TIN TỨC NỔI BẬT

1

Giới thiệu Open Agent Specification (Agent Spec): Một chuẩn 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 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 trên nhiều nền tảng và ứng dụng 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 tích hợp các tác nhân AI một cách liền mạch và tương tác hơn. Agent Spec dựa trên một mô hình dữ liệu đơn giản, dựa trên JSON, bắt giữ các đặc điểm thiết yếu của một tác nhân AI, bao gồm khả năng, ý định và hành vi của nó. Biểu diễn tiêu chuẩn hóa này cho phép các tác nhân AI giao tiếp và cộng tác với nhau, cũng như với các hệ thống và ứng dụng khác, một cách hiệu quả và hiệu lực hơn. Bằng cách áp dụng Agent Spec, các nhà phát triển có thể tạo ra các tác nhân AI linh hoạt, thích nghi và có thể mở rộng hơn, và có thể dễ dàng tích hợp vào nhiều ứng dụng và dịch vụ khác nhau. Agent Spec là một bước quan trọng hướng tới việc tạo ra một hệ sinh thái AI mở và kết nối hơn, nơi các tác nhân AI có thể làm việc cùng nhau để đạt được các mục tiêu chung và cung cấp kết quả tốt hơn cho người dùng.

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 Bedrock AgentCore MCP là một thành phần chính của nền tảng Amazon Bedrock, cho phép các nhà phát triển xây dựng, đào tạo và triển khai các model AI với quy mô lớn. Máy chủ AgentCore MCP là một máy chủ có khả năng mở rộng cao và bảo mật, cung cấp nền tảng cho việc xây dựng và đào tạo các model AI. Nó được tối ưu hóa cho hiệu suất và có thể xử lý lượng dữ liệu lớn, khiến nó trở nên lý tưởng cho các công việc AI phức tạp. Máy chủ này cũng hỗ trợ nhiều framework và công cụ AI, cho phép các nhà phát triển chọn cách tiếp cận tốt nhất cho nhu cầu cụ thể của họ. Bằng cách sử dụng máy chủ Amazon Bedrock AgentCore MCP, các nhà phát triển có thể tăng tốc quá trình phát triển AI, giảm chi phí và cải thiện hiệu quả tổng thể của các quy trình làm việc. Máy chủ này là một phần của nền tảng Amazon Bedrock, cung cấp một bộ công cụ và dịch vụ toàn diện cho việc xây dựng và triển khai các model AI.

3 Agent Factory: Từ prototype đến production—các công cụ dành cho developer 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 bộ công cụ dành cho developer được thiết kế để thúc đẩy sự phát triển và triển khai nhanh chóng các agent. Nền tảng này cho phép người dùng tạo, thử nghiệm và triển khai các agent thông minh trên nhiều môi trường khác nhau, bao gồm Azure. Agent Factory tối ưu hóa quá trình phát triển bằng cách cung cấp một loạt các công cụ và tính năng giúp đơn giản hóa việc tạo ra các hệ thống dựa trên agent phức tạp. Với Agent Factory, các developer có thể nhanh chóng xây dựng và triển khai các agent có thể tương tác với nhiều nguồn dữ liệu, dịch vụ và ứng dụng khác nhau. Nền tảng này hỗ trợ nhiều ngôn ngữ lập trình và tích hợp liền mạch với các dịch vụ Azure, cho phép người dùng tận dụng tối đa tiềm năng của đám mây. Bằng cách cung cấp một bộ công cụ toàn diện và giao diện người dùng thân thiện, Agent Factory nhằm mục đích tăng tốc sự phát triển và triển khai các agent thông minh, cho phép người dùng tập trung vào xây dựng các giải pháp sáng tạo thay vì quản lý cơ sở hạ

tầng phức tạp. Agent Factory là một phần của Microsoft Azure, cho phép người dùng tận dụng khả năng mở rộng và độ tin cậy của đám mây để triển khai và quản lý các agent của họ.

4

3 startup hạt nhân đạt cột mốc lớn. Tại sao điều đó quan trọng—và tại sao không.

3 Nuclear Startups Hit a Big Milestone. Why It Matters—and Why It Doesn't

Wired Đọc bài viết →

Ba công ty khởi nghiệp hạt nhân, Valar Atomics, Antares Nuclear và Deployable Energy, đã đạt được một cột mốc quan trọng bằng cách bật các lò phản ứng mới trong khuôn khổ một chương trình thí điểm do Bộ Năng lượng khởi xướng. Đây là một bước tiến hướng tới "phục hưng hạt nhân của Mỹ", như Bộ trưởng Năng lượng Chris Wright gọi nó, nhằm phát triển và triển khai thế hệ năng lượng nguyên tử tiếp theo. Chương trình thí điểm này được đưa vào hoạt động bởi một lệnh hành pháp từ Tổng thống Donald Trump, nhằm mục đích có ít nhất ba lò phản ứng đạt tới độ tới hạn trước ngày 4 tháng 7. Mặc dù thành tựu này được coi là một bước phát triển tích cực cho ngành công nghiệp, nhưng các chuyên gia cảnh báo rằng vẫn còn một chặng đường dài trước khi các thiết kế lò phản ứng mới trở thành hiện thực thương mại. Các công ty khởi nghiệp này đã được hưởng lợi từ sự hỗ trợ của chính phủ, bao gồm cả sự giúp đỡ từ các phòng thí nghiệm quốc gia được tài trợ bởi liên bang và các cắt giảm quy định đã đơn giản hóa quá trình. Tuy nhiên, các chuyên gia lưu ý rằng những nguyên mẫu này vẫn chưa phải là sản phẩm thương mại, mà chỉ là các lò phản ứng thử nghiệm.

5

Bộ thuật ngữ AI duy nhất bạn cần trong năm nay

The only AI glossary you'll need this year

TechCrunch AI Đọc bài viết →

Trong lĩnh vực trí tuệ nhân tạo đang phát triển nhanh chóng, một ngôn ngữ mới đã xuất hiện để mô tả khả năng và ứng dụng của nó. Để giúp điều hướng cảnh quan phức tạp này, một từ điển toàn diện đã được tạo ra để định nghĩa các thuật ngữ chính. Trí tuệ nhân tạo tổng quát (AGI) để cập đến AI vượt qua khả năng của con người trong các nhiệm vụ khác nhau, với định nghĩa khác nhau giữa các chuyên gia. Một tác nhân AI là một công cụ sử dụng các công nghệ AI để thực hiện các nhiệm vụ thay mặt cho người dùng, chẳng hạn như lập hóa

đơn hoặc viết mã. Cơ sở hạ tầng đang được xây dựng để hỗ trợ các tác nhân AI, có thể rút ra từ nhiều hệ thống AI để thực hiện các nhiệm vụ phức tạp. Các điểm cuối API, hoặc "nút" ở mặt sau của phần mềm, cho phép các chương trình tương tác với nhau, và các tác nhân AI ngày càng có thể tìm và sử dụng các điểm cuối này một cách tự động. Lý luận chuỗi suy nghĩ là một kỹ thuật được sử dụng bởi các mô hình ngôn ngữ lớn (LLM) để chia nhỏ các vấn đề phức tạp thành các bước nhỏ hơn, cải thiện độ chính xác của câu trả lời cuối cùng. Quá trình này liên quan đến việc phát triển các mô hình suy luận từ các mô hình ngôn ngữ truyền thống và tối ưu hóa chúng cho suy nghĩ chuỗi suy nghĩ.

6

Hóa đơn coding agent của bạn tăng gấp đôi. Đây là cách khắc phục.

Your coding agent bill doubled. Here's how to fix it.

LangChain Blog Đọc bài viết →

Chi phí cho các tác nhân mã hóa đã tăng vọt đối với nhiều công ty vào năm 2026, với một số nhóm trải qua sự tăng gấp sáu lần chỉ trong hai quý. Sự tăng đột ngột này không phải do sự tăng đáng kể về công việc, mà là do thiếu giám sát và quản lý. Kết quả là, các nhóm đang gặp khó khăn trong việc mở rộng công việc AI của họ đồng thời giữ chi phí dưới sự kiểm soát. Để giải quyết vấn đề này, các nhóm cần một cách đáng tin cậy để theo dõi và đo lường hiệu suất của các tác nhân mã hóa của họ, hiện đang ghi nhật ký hoạt động ở các định dạng khác nhau trên các công cụ khác nhau. Để đạt được một cái nhìn thống nhất, các nhóm có thể sử dụng một giải pháp tích hợp với nhiều tác nhân, chẳng hạn như Engine và LLM Gateway, để cung cấp một cái nhìn nhất quán và duy nhất về các phiên mã hóa của họ. Điều này cho phép các nhóm xác định các phiên tốn kém, các cuộc gọi công cụ thất bại và so sánh hành vi trên các tác nhân khác nhau. Tuy nhiên, việc thiết lập khác nhau cho mỗi công cụ và các nhóm phải tuân theo các bước cụ thể để tích hợp. Bằng cách có cái nhìn rõ ràng về việc sử dụng tác nhân mã hóa của họ, các nhóm có thể tối ưu hóa chi tiêu, thiết lập quản trị và bảo vệ lợi ích của họ, cuối cùng dẫn đến các công việc AI hiệu quả hơn.

7

Giải mã loop engineering: Khai thác nhiều hơn từ các AI agent, tránh "loopmaxxing"

Demystifying loop engineering: Get more from AI agents, avoid loopmaxxing

BD Tech Talks Đọc bài viết →

Trong lĩnh vực trí tuệ nhân tạo (AI), một sự thay đổi đáng kể đang diễn ra trong cách các nhà phát triển xây dựng ứng dụng với các mô hình ngôn ngữ lớn (LLM). Thay vì trực tiếp yêu cầu các tác nhân mã hóa, các nhà phát triển hiện đang tập trung vào việc thiết kế các vòng lặp yêu cầu các tác nhân đó. Sự thay đổi này là một phần của xu hướng rộng lớn hơn, với các đội kỹ sư AI chuyển từ việc yêu cầu trực tiếp mô hình và chuyển sang viết các vòng lặp thực hiện bên ngoài để phối hợp các hành động của mô hình. Sự thay đổi này, được gọi là kỹ thuật vòng lặp, đại diện cho một sự thay đổi□□ chức năng, nơi các nhà phát triển tập trung vào việc tạo ra dây chuyền lắp ráp phần mềm rộng lớn hơn, cho phép hệ thống đánh giá các đầu ra trung gian và tự xác định các bước tiếp theo một cách tự chủ. Tuy nhiên, các vòng lặp được thiết kế kém có thể dẫn đến các vấn đề như chi phí tăng vọt, giảm khả năng quan sát và theo đuổi mục tiêu không hiệu quả. Để xây dựng các vòng lặp hiệu quả, các nhà phát triển cần xem xét các nguyên tắc cấu trúc cụ thể, bao gồm theo dõi trạng thái bền vững, các plugin bên ngoài và các rào cản hoạt động. Một vòng lặp sẵn sàng sản xuất yêu cầu một kích hoạt hoạt động riêng biệt và một điều kiện thoát có thể xác minh. Sự tiến hóa này trong phát triển AI dự kiến sẽ tiếp tục, với các nền tảng phát triển chính sản phẩm hóa các mẫu này và nhúng các lệnh "mục tiêu" và "vòng lặp" trực tiếp vào các công cụ, cho phép các nhà phát triển tương tác với API và framework một cách hiệu quả hơn.

8

Bảo tàng Toàn cầu của IEEE mang lịch sử kỹ thuật đến với bạn

IEEE's Global Museum Brings Engineering History to You

IEEE Spectrum Đọc bài viết →

Bảo tàng Toàn cầu IEEE là một chương trình triển lãm lưu động giới thiệu lịch sử kỹ thuật và công nghệ đến công chúng. Bảo tàng được tạo ra bởi nhóm Lịch sử và Di sản IEEE để trưng bày các hiện vật kỹ thuật lịch sử được các thành viên IEEE quyên góp. Các triển lãm nhằm mục đích giáo dục mọi người về sự tiến bộ của các bước tiến công nghệ qua các thế hệ và cách các kỹ sư xây dựng trên những thành tựu trong quá khứ để mang lại lợi ích cho nhân loại. Bảo tàng đã□□ một số triển lãm, bao gồm "Tín hiệu không nhìn thấy: Cuộc cách mạng Radio của E. Howard Armstrong", nổi bật cuộc đời và tác động của nhà phát minh Edwin Howard Armstrong đối với ngành công nghiệp phát thanh và truyền thông không dây. Triển lãm này trưng bày một mẫu radio

> siêu heterodyne hiếm và các hiện vật khác, bao gồm cả Audion và Motorola Walkie-Talkie. Bảo tàng cũng có các triển lãm về vi mạch, công nghệ radio và các chủ đề khác. Ví dụ, triển lãm "Vi mạch làm thay đổi thế giới" khám phá vai trò của mạch tích hợp trong các lĩnh vực khác nhau, bao gồm xử lý tín hiệu và viễn thông. Bảo tàng Toàn cầu IEEE hợp tác với các hội IEEE để kỷ niệm các dịp lễ và đã hợp tác với các địa điểm khác nhau, bao gồm thư viện, trường đại học và bảo tàng, để trưng bày các triển lãm của mình.

TIPS & TRICKS CHO DEV

Tối ưu hóa Retrieval-Augmented Generation

Vấn đề: Hiệu suất của mô hình RAG giảm do không tối ưu hóa được dữ liệu đầu vào.
Cách làm: Sử dụng kỹ thuật fine-tuning để điều chỉnh mô hình cho phù hợp với dữ liệu cụ thể, ví dụ với lệnh `transformers.Trainer` trong thư viện Hugging Face.
Đánh giá: Hiệu quả cao khi áp dụng cho các mô hình lớn, nhưng cần phải có dữ liệu chất lượng cao.

Tìm kiếm ngữ nghĩa với embeddings

Vấn đề: Kết quả tìm kiếm không chính xác do không sử dụng thông tin ngữ nghĩa.
Cách làm: Sử dụng embeddings như Word2Vec hoặc BERT để mã hóa từ và câu, sau đó áp dụng kỹ thuật tính tương tự như cosine similarity, ví dụ `sentence-transformers` trong Python.
Đánh giá: Mang lại kết quả tìm kiếm chính xác hơn, đặc biệt trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Áp dụng semantic search cho dữ liệu lớn

Vấn đề: Tìm kiếm thông tin trong dữ liệu lớn trở nên khó khăn và tốn thời gian.
Cách làm: Sử dụng thư viện như `faiss` hoặc `pinecone` để triển khai tìm kiếm ngữ nghĩa trên dữ liệu lớn, ví dụ với lệnh `faiss.index_flat_l2` để tạo chỉ số tìm kiếm.
Đánh giá: Hiệu quả cao khi áp dụng cho các cơ sở dữ liệu lớn, giúp giảm thời gian tìm kiếm và tăng độ chính xác.

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). Điều này giúp giảm thiểu chi phí vận hành và cải thiện trải nghiệm người dùng. Việc tối ưu hóa LLM cũng cho phép các dev

triển khai ứng dụng trên các thiết bị có tài nguyên hạn chế.
3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và cải thiện hiệu suất.
4. Tip hoặc bước tiếp theo: Các dev nên nghiên cứu và áp dụng các kỹ thuật như quantization, pruning và knowledge distillation để tối ưu hóa LLM và cải thiện hiệu suất của ứng dụng AI.