BIS - MT · 08/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Có công mài sắt, có ngày nên kim.”

— Tục ngữ Việt Nam

Kiên trì mài giũa dù việc nhỏ nhất — mọi thành tựu đều đến từ nỗ lực bền bỉ theo thời gian.

TIN TỨC NỔI BẬT

1

Giới thiệu Open Agent Specification (Agent Spec): Định dạng 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 chung cho các nhà phát triển để tạo và tích hợp 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 xây dựng các tác nhân AI có thể tương tác với nhau một cách liền mạch. Agent Spec định nghĩa một tập hợp các giao diện và cấu trúc dữ liệu tiêu chuẩn cho các tác nhân AI, cho phép chúng chia sẻ kiến thức, phối hợp hành động và giao tiếp với nhau. Biểu diễn thống nhất này dự kiến sẽ thúc đẩy sự phát triển của các hệ thống AI tinh vi và tự chủ hơn. Bằng cách áp dụng Agent Spec, các nhà phát triển có thể tạo ra các tác nhân AI tương thích với nhau, bất kể công nghệ hoặc nền tảng cơ bản. Tính tương tác này rất quan trọng cho việc áp dụng rộng rãi AI trong các ngành công nghiệp khác nhau, bao gồm chăm sóc sức khỏe, tài chính và giao thông. Agent Spec là một bước quan trọng hướng tới tạo ra một hệ sinh thái kết nối và thông minh hơn.

2

Agent Factory: Từ prototype đến production—các công cụ 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 Azure đã giới thiệu Agent Factory, một công cụ mới dành cho developer được thiết kế để tối ưu hóa quá trình tạo và triển khai các tác nhân thông minh. Agent Factory cho phép developer xây dựng, thử nghiệm và triển khai các tác nhân có thể thực hiện các nhiệm vụ phức tạp, chẳng hạn như phân tích dữ liệu và ra quyết định. Công cụ này được xây dựng trên cơ sở hạ tầng đám mây của Azure, cung cấp một môi trường có thể mở rộng và bảo mật cho việc phát triển tác nhân. Với Agent Factory, developer có thể tạo tác nhân bằng nhiều ngôn ngữ lập trình và framework khác nhau, bao gồm Python và C#. Công cụ này cũng bao gồm một loạt các mẫu và thư viện đã được xây dựng sẵn để giúp đẩy nhanh quá trình phát triển. Ngoài ra, Agent Factory cung cấp một giao diện thân thiện với người dùng để thử nghiệm và triển khai tác nhân, giúp developer dễ dàng đưa ý tưởng của mình vào cuộc sống. Bằng cách tận dụng Agent Factory, developer có thể tạo ra các tác nhân thông minh có thể tự động hóa các nhiệm vụ, nâng cao trải nghiệm của khách hàng và thúc đẩy thông tin kinh doanh. Công cụ này được thiết kế để hỗ trợ việc phát triển tác nhân nhanh chóng, cho phép developer phản ứng nhanh với nhu cầu kinh doanh thay đổi và vượt qua đối thủ cạnh tranh.

3

Tăng tốc phát triển với server Amazon Bedrock AgentCore MCP | AI

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 này là một phần của nền tảng Amazon Bedrock, cung cấp một môi trường toàn diện để xây dựng, đào tạo và triển khai các model AI. Máy chủ Amazon Bedrock AgentCore MCP là một máy chủ hiệu suất cao cho phép các nhà phát triển đào tạo và triển khai các model AI một cách nhanh chóng và hiệu quả. Nó được tối ưu hóa cho các khối lượng công việc học máy và cung cấp một môi trường có thể mở rộng và bảo mật cho phát triển AI. Với máy chủ Amazon Bedrock AgentCore MCP, các nhà phát triển có thể tận dụng các khả năng AI tiên tiến của AWS, bao gồm hỗ trợ cho các framework và công cụ học sâu phổ biến. Điều này cho phép phát triển và triển khai các model AI nhanh hơn, giúp các tổ chức phản ứng nhanh với nhu cầu kinh doanh thay đổi và duy trì lợi thế cạnh tranh. Máy chủ

Amazon Bedrock AgentCore MCP là một thành phần chính của nền tảng Amazon Bedrock, cung cấp một môi trường mạnh mẽ và bảo mật cho phát triển và triển khai AI.

4

Cập nhật Innovation Graph Q1 2026: Hợp tác open source đang tăng tốc trên toàn cầu

Q1 2026 Innovation Graph update: Open source collaboration is accelerating worldwide

GitHub Blog Đọc bài viết →

Cập nhật Biểu đồ Đổi mới (Innovation Graph) Q1 2026 mới nhất từ GitHub cho thấy sự tăng tốc đáng kể trong hợp tác mã nguồn mở trên toàn thế giới. Dữ liệu mới cho thấy các cộng đồng developer toàn cầu đang phát triển nhanh hơn bao giờ hết, với hợp tác đạt mức cao mới trên các nền kinh tế khác nhau. Chỉ số cộng tác kinh tế (economy collaborators) của Biểu đồ Đổi mới GitHub, đo lường hợp tác xuất cảnh, đã trải qua tốc độ tăng trưởng 16% theo quý từ Q4 2025 đến Q1 2026. Đây là tốc độ tăng trưởng theo quý cao thứ hai được thấy kể từ năm 2020, với mức cao nhất là 21% trong Q2 2020. Sự tăng trưởng trong hợp tác mã nguồn mở là minh chứng cho việc áp dụng ngày càng nhiều phương pháp mã nguồn mở trong phát triển phần mềm. Xu hướng này dự kiến sẽ tiếp tục, với các tổ chức trên toàn thế giới tích hợp các nguyên tắc mã nguồn mở vào quy trình phát triển của họ. Cập nhật này cũng nhấn mạnh tầm quan trọng của hợp tác và đổi mới trong hệ sinh thái developer, với GitHub được vị trí là một nhà lãnh đạo trong ngành.

5

Cuộc chạy đua AI đã chuyển dịch từ các model thông minh sang hạ tầng full-stack như thế nào

How the AI arms race moved from smart models to full-stack infrastructure

BD Tech Talks Đọc bài viết →

Ngành công nghiệp trí tuệ nhân tạo (AI) đã chuyển từ tập trung vào việc phát triển các mô hình thông minh sang một cuộc chiến cơ sở hạ tầng rộng lớn hơn. Bộ xếp chồng AI hiện đại bao gồm bốn lớp chính: phần cứng, cụm tính toán, lớp mô hình và lớp ứng dụng. Các công ty không còn có thể dựa vào việc thống trị một lớp duy nhất để duy trì lợi thế cạnh tranh. Các công ty thành công đang mở rộng sang các lớp liền kề để bảo vệ biên độ và nắm bắt giá trị mới. Để mở rộng quy mô hoạt động, các công ty đang đầu tư vào cơ sở hạ tầng tùy chỉnh, chẳng

hạn như chip suy luận AI tùy chỉnh và trung tâm dữ liệu đa gigawatt. OpenAI đang hợp tác với Broadcom để ra mắt chip suy luận AI tùy chỉnh đầu tiên, trong khi Anthropic đã ký kết một thỏa thuận cơ sở hạ tầng 50 tỷ USD với nhà cung cấp neocloud FluidStack. xAI đã áp dụng một cách tiếp cận khác bằng cách xây dựng siêu máy tính của riêng mình, Colossus, và cho thuê khả năng tính toán cho các công ty khác. Trong khi đó, các công ty lớp ứng dụng đang phát triển tích hợp công việc sâu và tích lũy dữ liệu người dùng để tạo ra các rào cản phòng thủ. Cursor, một trợ lý mã hóa, đã xây dựng một mô hình tinh chỉnh, Composer 2.5, bằng cách sử dụng dữ liệu UX của riêng mình, giảm sự phụ thuộc vào các API tiền phong tốn kém. Điều này đã khiến các công ty lớp ứng dụng trở nên hấp dẫn đối với các gã khổng lồ tính toán lớp cơ sở, với SpaceX mua lại nhà phát triển Anysphere của Cursor trong một thỏa thuận 60 tỷ USD.

6

Meta giờ đây cho phép bất kỳ ai sử dụng ảnh Instagram của bạn trong các ảnh AI—Trừ khi bạn từ chối

Meta Now Lets Anyone Use Your Instagram Photos in AI Images—Unless You Opt Out

Wired Đọc bài viết →

Meta đã ra mắt mô hình hình ảnh AI của mình, Muse Image, cho phép người dùng tạo hình ảnh bằng cách sử dụng ngoại hình của những người thực. Trong khuôn khổ cập nhật này, các hồ sơ Instagram công khai sẽ tự động được chọn để sử dụng cho các bản remix AI tạo sinh. Người dùng có thể gắn thẻ hồ sơ của một tài khoản công khai trong một lời nhắc để sử dụng ngoại hình của họ trong một hình ảnh. Meta định vị tính năng này như một cách để cá nhân hóa hình ảnh, nhưng người dùng có thể chọn không tham gia bằng cách điều chỉnh cài đặt của họ. Để thực hiện việc này, người dùng phải mở ứng dụng Instagram, chạm vào hồ sơ của họ và điều hướng đến tab Chia sẻ và tái sử dụng, nơi họ có thể tắt tùy chọn cho phép mọi người sử dụng nội dung của họ trên Instagram và với các tính năng AI trên Meta. Thay đổi này sẽ ngăn chặn việc tạo ra các hình ảnh bổ sung, nhưng các hình ảnh AI hiện có được tạo bằng nội dung của người dùng sẽ không bị xóa.

7

sqlite-utils 4.0, giờ đây với các database schema migration

sqlite-utils 4.0, now with database schema migrations

Simon Willison Đọc bài viết →

Một phiên bản mới của thư viện sqlite-utils đã được phát hành, phiên bản 4.0. Bản cập nhật lớn này bao gồm ba tính năng quan trọng: di chuyển cơ sở dữ liệu, giao dịch lồng nhau và hỗ trợ khóa ngoại hợp chất. Di chuyển cơ sở dữ liệu cho phép các developer định nghĩa một chuỗi các thay đổi sẽ được thực hiện trên cơ sở dữ liệu SQLite, theo dõi những di chuyển nào đã được áp dụng và áp dụng bất kỳ di chuyển nào đang chờ xử lý. Điều này được thực hiện thông qua các tệp Python sử dụng thư viện sqlite-utils và phương thức table.transform() mạnh mẽ của nó, giúp tăng cường khả năng của câu lệnh ALTER TABLE của SQLite. Bảng _sqlite_migrations được sử dụng để theo dõi các hàm di chuyển đã được áp dụng. Người dùng có thể tạo tệp di chuyển, chẳng hạn như migrations.py, và chạy chúng chống lại cơ sở dữ liệu bằng lệnh sqlite-utils migrate. Di chuyển cũng có thể được thực hiện từ mã Python bằng phương thức migrations.apply(). Tính năng này được thiết kế để đơn giản và linh hoạt, cho phép các developer quản lý lược đồ cơ sở dữ liệu của họ trên nhiều phiên bản. Thư viện sqlite-utils đã được cập nhật để bao gồm tính năng này, tính năng trước đây có sẵn trong một gói riêng biệt gọi là sqlite-migrate.

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 giới thiệu lịch sử kỹ thuật và tiến bộ công nghệ thông qua các triển lãm lưu động. Bảo tàng được tạo ra bởi nhóm Lịch sử và Di sản IEEE để trưng bày các hiện vật kỹ thuật lịch sử được thu thập bởi các thành viên IEEE. Các triển lãm này được thiết kế để giáo dục công chúng về cách các tiến bộ công nghệ đã diễn ra qua các thế hệ và cách các kỹ sư xây dựng trên những thành tựu trong quá khứ. Triển lãm chủ chốt của bảo tàng, "Tín hiệu không nhìn thấy: Cách mạng Radio của E. Howard Armstrong," trưng bày một nguyên mẫu radio hiếm được phát triển bởi Edwin Howard Armstrong, người đã phát minh ra hệ thống radio FM. Triển lãm cũng bao gồm các hiện vật khác, chẳng hạn như Audion được sử dụng trong các thí nghiệm của Armstrong và một Motorola Walkie-Talkie từ Chiến tranh Triều Tiên. Các triển lãm khác đã tập trung vào các chủ đề như vi mạch, phát minh radio sớm, và công nghệ điện và truyền thông. Bảo tàng hợp tác với các hội IEEE để tạo ra các triển lãm đánh dấu các dịp kỷ niệm và cột mốc quan trọng trong lĩnh vực kỹ thuật. Các triển lãm đã được

> công chúng đón nhận nồng nhiệt, với nhiều người tham dự thể hiện sự kết nối cảm xúc với các hiện vật và câu chuyện được trưng bày.

TIPS & TRICKS CHO DEV

Tối ưu hóa context window

Vấn đề: Context window quá nhỏ gây hạn chế khả năng xử lý văn bản dài.
Cách làm: Sử dụng kỹ thuật chunking, chia văn bản thành đoạn nhỏ hơn. Ví dụ, sử dụng lệnh `--max_length` trong CLI để giới hạn độ dài văn bản.
Đánh giá: Hiệu quả khi xử lý văn bản dài, nhưng có thể mất ngữ nghĩa nếu không được thực hiện đúng.

Quản lý long-context

Vấn đề: Long-context quá lớn gây tốn tài nguyên và giảm hiệu suất.
Cách làm: Sử dụng kỹ thuật phân đoạn, chia long-context thành nhiều phần nhỏ hơn. Ví dụ, sử dụng prompt "Summarize the following text" để giảm độ dài văn bản.
Đánh giá: Hiệu quả khi xử lý văn bản dài và phức tạp, nhưng cần được thực hiện cẩn thận.

Tối ưu hóa memory

Vấn đề: Memory quá nhỏ gây hạn chế khả năng xử lý dữ liệu.
Cách làm: Sử dụng kỹ thuật caching, lưu trữ dữ liệu tạm thời để giảm tải bộ nhớ. Ví dụ, sử dụng lệnh `--cache_dir` trong CLI để chỉ định thư mục cache.
Đánh giá: Hiệu quả khi xử lý dữ liệu lớn, nhưng cần được cấu hình đúng để tránh mất dữ liệu.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
Dev cần biết về tối ưu chi phí và hiệu năng LLM để giảm thiểu chi phí vận hành và tăng tốc độ xử lý của các mô hình AI. Điều này giúp các ứng dụng AI trở nên hiệu quả và tiết kiệm hơn.

2. Việc tối ưu hóa LLM liên quan đến việc điều chỉnh các tham số và cấu hình của mô hình để đạt được hiệu suất tốt nhất.

3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm thiểu chi phí và tăng tốc độ xử lý của mô hình.

4. Tip: Để bắt đầu tối ưu hóa LLM, hãy bắt đầu bằng việc phân tích các tham số và cấu hình của mô hình hiện tại, sau đó thử nghiệm với các kỹ thuật như fine-tuning và LoRA để tìm ra cách tối ưu hóa hiệu suất và chi phí.