BIS - MT · 13/08/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“In the middle of every difficulty lies opportunity.”

↪ Giữa mọi khó khăn đều ẩn chứa một cơ hội.

— Albert Einstein

Thay vì chỉ thấy vấn đề, hãy tìm kiếm cơ hội ẩn bên trong — tư duy tích cực là nền tảng của sự sáng tạo.

TIN TỨC NỔI BẬT

1 Cách chạy các AI model cục bộ trên RTX 5060 Ti của bạn — Hướng dẫn từng bước

How to Run Local AI Models on Your RTX 5060 Ti — Step-by-Step Guide

TechnoSports Media Group Đọc bài viết →

> Chạy các mô hình AI cục bộ trên máy tính cá nhân có thể là một giải pháp hiệu quả và tiết kiệm chi phí so với các dịch vụ dựa trên đám mây. Card đồ họa NVIDIA RTX 5060 Ti là một công cụ mạnh mẽ cho mục đích này. Dưới đây là hướng dẫn từng bước để giúp bạn chạy các mô hình AI cục bộ trên RTX 5060 Ti của mình. Trước tiên, hãy đảm bảo hệ thống của bạn đáp ứng các yêu cầu tối thiểu, bao gồm card đồ họa RTX 5060 Ti, CPU tương thích và RAM đủ. Tiếp theo, cài đặt phần mềm cần thiết, bao gồm NVIDIA CUDA Toolkit và một khuôn khổ học sâu như TensorFlow hoặc PyTorch. Một khi phần mềm đã được cài đặt, bạn có thể bắt đầu xây dựng và đào tạo các mô hình AI của mình bằng cách sử dụng khuôn khổ học sâu. Điều này liên quan đến việc viết mã, biên dịch mô hình và kiểm tra hiệu suất của nó. Để chạy mô hình cục bộ, sử dụng các công cụ tích hợp sẵn của khuôn khổ hoặc NVIDIA GPU để tăng tốc tính toán. Bằng cách làm theo các bước này, bạn có thể tận dụng sức mạnh của RTX 5060 Ti để chạy các mô hình AI cục bộ, đạt được thời gian xử lý nhanh hơn và hiệu suất tốt hơn.

2 Lớp lệnh AI kho hàng đa agent giúp tối ưu vận hành và nâng cao thông minh chuỗi cung ứng

Multi-Agent Warehouse AI Command Layer Enables Operational Excellence and Supply Chain Intelligence

NVIDIA Developer Đọc bài viết →

NVIDIA đã phát triển một Lớp lệnh AI Nhà kho Đa tác nhân, một công nghệ tiên tiến được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng trong các nhà kho. Giải pháp đổi mới này tận dụng trí tuệ nhân tạo (AI) để tối ưu hóa hoạt động nhà kho, cải thiện quản lý hàng tồn kho và tối ưu hóa hậu cần. Lớp lệnh AI Nhà kho Đa tác nhân cho phép theo dõi và phân tích thời gian thực các hoạt động nhà kho, cho phép đưa ra quyết định dựa trên dữ liệu. Bằng cách tích hợp các tác nhân được hỗ trợ bởi AI, hệ thống có thể dự đoán và ngăn chặn các nút thắt tiềm năng, tối ưu hóa phân bổ tài nguyên và cải thiện năng suất nhà kho tổng thể. Các lợi ích chính của công nghệ này bao gồm: - Hiệu quả hoạt động được nâng cao - Trí tuệ chuỗi cung ứng được cải thiện - Theo dõi và phân tích thời gian thực - Phân tích dự đoán cho việc ra quyết định chủ động - Phân bổ tài nguyên được tối ưu hóa Lớp lệnh AI Nhà kho Đa tác nhân của NVIDIA có tiềm năng cách mạng hóa cách thức hoạt động của các nhà kho, cho phép các doanh nghiệp duy trì lợi thế cạnh tranh và đáp ứng nhu cầu của một thị trường thay đổi nhanh chóng.

3 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 biểu diễn thống nhất cho các tác nhân AI. Agent Spec nhằm cung cấp một khuôn khổ tiêu chuẩn hóa để mô tả các tác nhân AI, cho phép khả năng tương tác và hợp tác 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 dễ dàng tích hợp và kết hợp các tác nhân AI từ các nguồn khác nhau. Agent Spec định nghĩa một cấu trúc chung cho các tác nhân AI, bao gồm khả năng, mục tiêu và hành vi của chúng. Cấu trúc này cho phép các tác nhân AI được mô tả trong một định dạng nhất quán và có thể đọc được bằng máy, giúp việc giao tiếp và tương tác giữa các hệ thống khác nhau trở nên dễ dàng. Bằng cách áp dụng Agent Spec, các nhà phát triển có thể tạo ra các ứng dụng AI phức tạp và tích hợp hơn, và các tổ chức có thể tận dụng một loạt các khả năng AI rộng lớn hơn. Việc giới thiệu Agent Spec dự kiến sẽ thúc

đẩy sự đổi mới và hợp tác trong cộng đồng AI, vì các nhà phát triển hiện có thể xây dựng trên một nền tảng chung và chia sẻ kiến thức và chuyên môn dễ dàng hơn.

4

DeepSeek V4 Pro 0813 (trên OpenRouter)

DeepSeek V4 Pro 0813 (on OpenRouter)

Simon Willison Đọc bài viết →

DeepSeek, một model trí tuệ nhân tạo, đã phát hành phiên bản mới nhất, DeepSeek V4 Pro 0813, có sẵn thông qua API OpenRouter. Tuy nhiên, không có trang thông báo chính thức cho model mới và trọng số của model vẫn chưa được xác nhận công khai. Hiệu suất của model có thể được tùy chỉnh thành ba mức lý luận khác nhau: thấp, trung bình và cao, dẫn đến đầu ra khác nhau, chẳng hạn như hình ảnh chim bồ câu khác nhau. Kết quả benchmark cho model có sẵn, nhưng ban đầu được chia sẻ trên Nhóm WeChat Chính thức của DeepSeek và sau đó được đăng trên Reddit và Hacker News. Việc phát hành model DeepSeek V4 Pro 0813 đánh dấu một bản cập nhật cho loạt DeepSeek V4, bao gồm các model trước đó như DeepSeek V4 Flash 0731.

5

Sau Orthogonality: Agency đạo đức đức hạnh và AI Alignment

After Orthogonality: Virtue-Ethical Agency and AI Alignment

The Gradient Đọc bài viết →

Bài viết này khám phá khái niệm về việc kết hợp trí tuệ nhân tạo (AI) với quyền lực và giá trị của con người. Tác giả cho rằng những người có lý trí không có mục tiêu, mà thay vào đó là điều chỉnh hành động của họ với các thực tiễn, những mạng lưới gồm các hành động, khuynh hướng và tiêu chí đánh giá. Để tạo ra các AI thực sự hỗ trợ quyền lực của con người, quá trình suy xét của các tác nhân AI phải chia sẻ một logic tương tự. Tác giả đề xuất rằng AI nên được thiết kế để thúc đẩy một số giá trị nhất định, chẳng hạn như minh bạch, hữu ích và vô hại, không phải là mục tiêu, mà là các động lực trong mạng lưới hành động. Cách tiếp cận này được lấy cảm hứng từ khái niệm về eudaimonia, hay sự thịnh vượng, hạnh phúc của con người một cách chủ động và hợp lý. Tác giả đề xuất rằng tính hợp lý eudaimonic, ưu tiên việc thúc đẩy giá trị trong các thực tiễn, là một hình thức đại diện tự nhiên và hiệu quả hơn cho quyền lực của con người so với tính hợp lý truyền thống theo hệ quả hoặc nghĩa vụ. Bài viết cho rằng việc kết

hợp AI với tính hợp lý eudaimonic có thể giúp giải quyết các mối quan ngại và nghịch lý an toàn AI cổ điển và rằng cách tiếp cận này phù hợp hơn với giá trị và đạo đức của con người.

6 Hàng terabyte credential bị rò rỉ trong cuộc tấn công chuỗi cung ứng quy mô lớn

Terabytes of credentials leaked in massive supply-chain attack

Ars Technica Đọc bài viết →

Một cuộc tấn công chuỗi cung ứng lớn đã lộ ra terabyte thông tin đăng nhập nhạy cảm từ hơn 2.500 tổ chức trên toàn thế giới. Cuộc tấn công nhắm vào LiteLLM, một công cụ mã nguồn mở được sử dụng cho phát triển phần mềm dựa trên AI, và các phiên bản bị xâm phạm của công cụ này đã được tải xuống từ kho lưu trữ Python Package Index chính thức. Các công ty bảo mật CloudSEK và Hudson Rock đã phát hiện ra sự vi phạm này, bao gồm các khóa cloud, token repository, khóa SSH và các thông tin nhạy cảm khác. Các phiên bản bị xâm phạm của LiteLLM cho phép kẻ tấn công truy cập vào bộ nhớ của máy bị nhiễm, thu thập dữ liệu và chuyển nó ra ngoài qua một kênh được kiểm soát bởi kẻ tấn công. Các thông tin đăng nhập bị lộ bao gồm các bí mật truy cập từ các tổ chức lớn như Microsoft, Amazon, Cisco, Samsung và Salesforce. Các nhà nghiên cứu đã kêu gọi các tổ chức bị ảnh hưởng phải thay đổi tất cả các thông tin đăng nhập trong đường ống của họ và thực hiện kiểm toán ngay lập tức môi trường của họ để tìm kiếm các phiên bản bị xâm phạm của LiteLLM. Cuộc tấn công này nhấn mạnh tầm quan trọng của việc ưu tiên bảo mật AI và bảo mật DevOps để ngăn chặn các sự vi phạm như vậy.

7 Streamer Twitch giờ đây có thể từ chối dùng nội dung để train AI của Amazon

Twitch streamers can now opt out from training Amazon's AI

The Verge AI Đọc bài viết →

Các streamer trên Twitch hiện có tùy chọn từ chối cho phép nội dung của họ được sử dụng để đào tạo các model AI tạo sinh của Amazon. Thay đổi này cho phép người dùng kiểm soát cách các luồng, VOD, clip và thông tin kênh của họ được sử dụng bởi AI của Amazon. Nếu một streamer chọn từ chối, nội dung của họ sẽ không được sử dụng trong việc đào tạo các model AI của Amazon trong tương lai, những model

này tạo ra hoặc tổng hợp văn bản, âm thanh, hình ảnh hoặc video. Tuy nhiên, việc từ chối sẽ không ảnh hưởng đến chức năng của các tính năng hỗ trợ AI như phụ đề và công cụ an toàn. Tùy chọn "Đào tạo cho AI tạo sinh" có thể được tìm thấy trong cài đặt của Twitch dưới tab Bảo mật và Quyền riêng tư. Điều đáng chú ý là việc từ chối đào tạo AI tạo sinh không có nghĩa là Twitch và Amazon sẽ không sử dụng nội dung của streamer cho các mục đích khác, chẳng hạn như thúc đẩy sự phát triển của streamer và an toàn cộng đồng.

8

Những bức ảnh đẹp nhất về nhật thực toàn phần tháng Tám

The Best Photos of the Big August Solar Eclipse

Wired Đọc bài viết →

Một lần nguyệt thực toàn phần có thể nhìn thấy ở các phần của Bắc Mỹ, châu Âu và Tây Phi vào ngày 13 tháng 8. Sự kiện thiên văn hiếm gặp này, xảy ra lần đầu tiên trong hơn một thế kỷ ở Tây Ban Nha, mang lại một cảnh tượng độc đáo và đẹp như tranh vẽ khi vành nhật hoa của mặt trời phát sáng trên đường chân trời. Con đường toàn phần, kéo dài khoảng bốn giờ rưỡi, đi qua bán cầu bắc từ tây sang đông, đi qua các khu vực như Galicia, Asturias và quần đảo Baleares ở Tây Ban Nha. Người dân tập trung để quan sát nguyệt thực, bao gồm cả người hâm mộ Real Madrid đã xem qua kính đặc biệt trước một trận đấu. Các nhà khoa học từ NASA và các cơ quan vũ trụ khác đã tiến hành các thí nghiệm dọc theo con đường toàn phần, thu thập dữ liệu về động lực học của vành nhật hoa và ảnh hưởng của nguyệt thực đối với bầu khí quyển của Trái Đất. Sự kiện này cho phép các nhà nghiên cứu nghiên cứu vành nhật hoa một cách chi tiết, điều chỉ có thể thực hiện được trong các lần nguyệt thực toàn phần, nhưng không liên quan gì đến AI, API, LLM, model, token, developer, framework, vì vậy không có từ nào trong số đó được sử dụng trong bản dịch này.

TIPS & TRICKS CHO DEV

Tối ưu hóa Retrieval-Augmented Generation

Vấn đề: Hiệu suất thấp khi tạo văn bản dài.

Cách làm: Sử dụng RAG với kỹ thuật chunking, chia văn bản thành đoạn nhỏ. Ví dụ, prompt cho Claude: "Tóm tắt văn bản sau thành 5 đoạn".

Đánh giá: Cải thiện hiệu suất tạo văn bản, nhưng cần tinh chỉnh kỹ thuật chunking.

Tạo embeddings hiệu quả

Vấn đề: Embeddings không chính xác do dữ liệu thiếu đa dạng.

Cách làm: Sử dụng thư viện Hugging Face Transformers, tạo embeddings với lệnh `python -m transformers create_embedding`.

Đánh giá: Cải thiện độ chính xác, nhưng yêu cầu dữ liệu đa dạng và chất lượng.

Tích hợp semantic search

Vấn đề: Tìm kiếm không chính xác do không hiểu ngữ nghĩa.

Cách làm: Sử dụng thư viện Faiss, tích hợp semantic search với lệnh `python -m faiss indexFlatL2`.

Đánh giá: Cải thiện độ chính xác tìm kiếm, nhưng cần tinh chỉnh thông số kỹ thuật.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Dev cần biết về tối ưu chi phí và hiệu năng LLM để đảm bảo ứng dụng AI của họ hoạt động hiệu quả và tiết kiệm chi phí. Điều này đặc biệt quan trọng khi triển khai mô hình ngôn ngữ lớn (LLM) trong sản xuất. Việc tối ưu hóa có thể giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý.
3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning để điều chỉnh mô hình LLM cho phù hợp với use case cụ thể, giúp giảm thiểu số lượng tham số cần đào tạo và tăng tốc độ đào tạo.
4. Tip: Sử dụng kỹ thuật LoRA (Low-Rank Adaptation) để tinh chỉnh mô hình LLM mà không cần đào tạo lại toàn bộ mô hình, giúp tiết kiệm thời gian và tài nguyên.