

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Làm người phải biết lo xa.”

— Tục ngữ Việt Nam

Tư duy dài hạn và lập kế hoạch trước giúp ta tránh được nhiều rủi ro và nắm bắt cơ hội tốt hơn.

TIN TỨC NỔI BẬT

1

Cách chạy 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 →](#)

Hướng dẫn từng bước đã được tạo ra để giúp người dùng chạy các mô hình AI cục bộ trên card đồ họa NVIDIA RTX 5060 Ti của họ. Hướng dẫn này nhằm cung cấp một quy trình trực tiếp cho những người muốn sử dụng RTX 5060 Ti của họ để thực thi mô hình AI. Để bắt đầu, người dùng cần đảm bảo hệ thống của họ đáp ứng các yêu cầu tối thiểu, bao gồm một RTX 5060 Ti, một CPU tương thích và RAM đủ. Tiếp theo, họ phải cài đặt phần mềm cần thiết, bao gồm CUDA Toolkit và thư viện cuDNN. Một khi phần mềm được cài đặt, người dùng có thể tiếp tục cài đặt khuôn khổ AI đã chọn của họ, chẳng hạn như TensorFlow hoặc PyTorch. Sau đó, họ sẽ cần cấu hình biến môi trường và biên dịch mô hình AI của họ. Cuối cùng, người dùng có thể chạy mô hình AI của họ trên RTX 5060 Ti, tận dụng sức mạnh của card đồ họa để tăng tốc tính toán AI. Hướng dẫn cung cấp hướng dẫn chi tiết cho từng bước, giúp người dùng với các cấp độ kỹ năng kỹ thuật khác nhau có thể tiếp cận.

2

Giới thiệu Open Agent Specification (Agent Spec): Một chuẩn biểu diễn thống nhất cho 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. Đặc tả này nhằm cung cấp một khuôn khổ chung cho các nhà phát triển để tạo, tích hợp và tương tác với các tác nhân AI trên nhiều nền tảng và ứng dụng khác nhau. Agent Spec định nghĩa một cấu trúc tiêu chuẩn hóa cho các tác nhân AI, cho phép chúng chia sẻ kiến thức, kỹ năng và hành vi một cách nhất quán và tương tác. Điều này giúp tạo ra các hệ thống AI phức tạp và động hơn, cũng như tích hợp các tác nhân AI với các hệ thống và ứng dụng khác. Đặc tả này được thiết kế để mở và có thể mở rộng, cho phép các nhà phát triển đóng góp vào sự phát triển và mở rộng khả năng của nó. Bằng cách cung cấp một biểu diễn thống nhất cho các tác nhân AI, Agent Spec có tiềm năng tăng tốc phát triển và triển khai các ứng dụng được hỗ trợ bởi AI trên nhiều ngành công nghiệp khác nhau. Việc giới thiệu Agent Spec dự kiến sẽ thúc đẩy tính tương tác, hợp tác và đổi mới trong lĩnh vực AI, cho phép các nhà phát triển xây dựng các hệ thống AI tinh vi và hiệu quả hơn.

3

Agent Factory: Kết nối agent, app và dữ liệu bằng 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, ứng dụng và dữ liệu bằng cách sử dụng các tiêu chuẩn mở. Nền tảng này tận dụng Microsoft Cloud PC (MCP) và tiêu chuẩn Ứng dụng đến Ứng dụng (A2A) để tạo điều kiện cho sự tương tác liền mạch giữa các hệ thống khác nhau. Agent Factory cho phép các developer tạo ra các agent tùy chỉnh có thể tương tác với các ứng dụng và nguồn dữ liệu khác nhau, thúc đẩy tính tương tác và linh hoạt. Bằng cách tận dụng các tiêu chuẩn MCP và A2A, các developer có thể xây dựng các giải pháp hiệu quả và có khả năng mở rộng hơn, tích hợp với một loạt các hệ thống. Việc giới thiệu Agent Factory là một phần trong nỗ lực của Microsoft nhằm thúc đẩy các tiêu chuẩn mở và tính tương tác trong ngành công nghệ. Bằng cách cung cấp một framework chung để kết nối các agent, ứng dụng và dữ liệu, Agent Factory nhằm mục đích đơn giản hóa quá trình phát triển và cho phép các giải pháp sáng tạo hơn. Cách tiếp cận dựa trên tiêu chuẩn mở của nền tảng này dự kiến sẽ tạo điều kiện cho sự hợp tác và tích hợp lớn hơn trên các hệ thống và ứng dụng khác nhau.

4

Hàng ngàn server có thể bị backdoor do lỗi trong bộ điều khiển bo mạch chủ

Thousands of servers can be backdoored by exploiting buggy motherboard controllers

Ars Technica

[Đọc bài viết →](#)

Hàng nghìn máy chủ kết nối internet được bán bởi các nhà sản xuất lớn dễ bị tấn công backdoor từ xa do các lỗ hổng bảo mật quan trọng trong các bộ điều khiển quản lý bo mạch (BMCs). Các nhà nghiên cứu đã phát hiện hơn một tá lỗ hổng mới trong BMCs từ các công ty như HPE, Supermicro, Avocent, Huawei, Lenovo, Dell và các công ty khác. Những lỗ hổng này cho phép các hacker thực thi mã độc từ xa trên các bộ điều khiển và có thể lây nhiễm cho các máy chủ mà chúng quản lý. Các vấn đề này đã tồn tại từ ít nhất năm 2013, khi các nhà nghiên cứu đầu tiên cảnh báo về rủi ro của BMCs. Mặc dù có nỗ lực để sửa chữa các vấn đề, nhưng nhiều lỗ hổng vẫn còn hoạt động và một số đã bị các hacker khai thác. Một cuộc quét gần đây của các BMCs kết nối internet đã tìm thấy hơn 86.000 thiết bị[] lộ dịch vụ quản lý cho công

chúng, với hơn 54% chứa các lỗ hổng bảo mật quan trọng. Một cuộc quét nội bộ của các mạng doanh nghiệp đã tìm thấy gần 29% BMCs có lỗ hổng bảo mật quan trọng. Nhà nghiên cứu đứng sau việc phát hiện này đã cảnh báo rằng các lỗ hổng này là phổ biến và dễ bị khai thác, khiến nó trở thành một rủi ro bảo mật đáng kể cho các trung tâm dữ liệu.

5

Vượt xa ReAct: Xây dựng AI agent stack hiện đại cho các hệ sinh thái tool khổng lồ

Beyond ReAct: Building the modern AI agent stack for massive tool ecosystems

BD Tech Talks

[Đọc bài viết →](#)

Các nhà nghiên cứu tại Alibaba đã phát triển một framework gọi là SkillWeaver để giải quyết thách thức trong việc tích hợp các hệ sinh thái công cụ quy mô lớn với các hệ thống AI. Khi các hệ thống AI xử lý các quy trình công việc phức tạp, chúng yêu cầu nhiều công cụ và kỹ năng, dẫn đến khó khăn trong việc chọn đúng kỹ năng cho mỗi nhiệm vụ. Tiêu chuẩn ngành hiện tại, ReAct, không thể xử lý các thư viện quy mô doanh nghiệp, đạt được độ chính xác phân tích 0%. Để vượt qua điều này, SkillWeaver sử dụng một kỹ thuật mới gọi là Phân tích Kỹ năng (SAD), nó sử dụng một vòng lặp phản hồi để chọn các công cụ phù hợp và cung cấp chúng cho Model Ngôn ngữ Lớn (LLM). Trong các thí nghiệm, SkillWeaver đã dẫn đến sự giảm đáng kể lên đến 99,9% trong việc tiêu thụ token và cải thiện việc hoàn thành nhiệm vụ và độ chính xác. Framework này hữu ích cho các ứng dụng AI thực tế nơi các tác nhân phải tự động hóa các hệ sinh thái công cụ do cộng đồng đóng góp cho các nhiệm vụ nhiều bước. SkillWeaver bao gồm ba giai đoạn: Phân tích, Tìm kiếm và Tổng hợp, chúng hoạt động cùng nhau để chia nhỏ các lời nhắc người dùng phức tạp thành các kế hoạch có thể thực thi. Điều này giải quyết các thách thức về giới hạn ngữ cảnh, tiêu thụ token và phân loại ồn ào giữa các API có mô tả tương tự nhưng khác biệt về lược đồ.

6

Đề xuất điều tra dân số mới: Không tính người nhập cư không giấy tờ, bỏ qua chủng tộc và xu hướng tính dục

Census Proposal Would Stop Counting Undocumented Immigrants—and Ignore Race and Sexual Orientation

Wired

[Đọc bài viết →](#)

Bộ Thương mại Hoa Kỳ đang đề xuất một quy định sẽ thay đổi đáng kể cuộc điều tra dân số năm 2030. Theo các tài liệu do WIRED thu được, quy định đề xuất sẽ ngừng thu thập dữ liệu dân số về các nhóm thiểu số chủng tộc và cá nhân LGBTQ+ và ngăn chặn việc đếm những người nhập cư không có giấy tờ. Quyết định này có thể có những tác động sâu rộng đối với tài trợ liên bang, đặc biệt là ở các bang có dân số thiểu số và cộng đồng nhập cư đáng kể. Quy định đề xuất cho rằng việc bao gồm những nhóm này sẽ làm sai lệch dữ liệu, nhưng các nhà phê bình lập luận rằng điều này sẽ khiến việc đánh giá sự bất bình đẳng và thách thức chính trị trở nên khó khăn. Quy định cũng có khả năng yêu cầu việc tái giới thiệu câu hỏi về quyền công dân, điều đã bị Tối cao Pháp viện bác bỏ trước đó. Các chuyên gia cảnh báo rằng sự thay đổi này có thể dẫn đến việc các địa phương có số lượng người nhập cư không có giấy tờ cao mất đi tài trợ liên bang quan trọng và ảnh hưởng đến dân số bản địa phụ thuộc vào tài trợ chuyên dụng. Bộ Thương mại chưa bình luận về đề xuất này.

7

Cách đánh giá Sourcegraph trên codebase của riêng bạn

How to evaluate Sourcegraph on your own codebase

Sourcegraph Blog

[Đọc bài viết →](#)

Để đánh giá Sourcegraph trên cơ sở mã của riêng bạn, cần có một cách tiếp cận tinh tế để đo lường tác động của nó đối với việc hoàn thành nhiệm vụ, tìm kiếm và chi phí. Hiệu quả của Sourcegraph phụ thuộc vào ý nghĩa của "tốt hơn" đối với nhóm của bạn, chẳng hạn như hoàn thành nhiều nhiệm vụ hơn, tìm kiếm mã nhanh hơn, giảm chi phí hoặc làm cho công việc khó khăn trở nên đáng tin cậy hơn. Một điểm chuẩn công khai không thể trả lời trực tiếp câu hỏi này, vì nó phụ thuộc vào cách công việc liên quan được phân phối trên cơ sở mã. Để đánh giá chính xác Sourcegraph, bạn nên đo lường riêng việc tìm kiếm, hiệu suất của agent và chi phí, vì chúng đo lường các phần khác nhau của hệ thống. Một phương pháp để chạy so sánh này bao gồm phân loại nhiệm vụ theo độ khó định vị và loại bỏ các nhiệm vụ đã cung cấp cho agent thông tin cần thiết. Cách tiếp cận này sẽ giúp bạn xác định liệu Sourcegraph có cải thiện việc tìm kiếm, hoàn thành nhiệm vụ và chi phí trong trường hợp sử dụng cụ thể của bạn hay không.

Ứng phó với những năng lực cyber quan trọng thế hệ mới

Responding to the next frontier of critical cyber capabilities

OpenAI Blog

[Đọc bài viết →](#)

OpenAI đang thực hiện các bước để nâng cao bảo mật cho công nghệ Astra của mình, một khả năng mạng quan trọng. Trong khuôn khổ nỗ lực này, công ty đang phát hành các đánh giá sơ bộ về an ninh mạng cho Astra. Động thái này nhằm cung cấp minh bạch và cái nhìn sâu sắc về các biện pháp được thực hiện để tăng cường các biện pháp bảo vệ và kiểm soát an ninh. Bằng cách chia sẻ những đánh giá này, OpenAI đang thể hiện cam kết của mình trong việc đảm bảo tính toàn vẹn và bảo mật của công nghệ của mình. Những đánh giá này có thể sẽ chi tiết cách tiếp cận của công ty trong việc xác định và giảm thiểu các điểm yếu tiềm năng, cũng như việc triển khai các biện pháp kiểm soát an ninh mạnh mẽ để ngăn chặn truy cập hoặc lạm dụng không được phép. Sự phát triển này rất quan trọng, vì nó cho thấy OpenAI đang ưu tiên bảo mật và độ tin cậy của công nghệ của mình, đặc biệt là trong các lĩnh vực được coi là quan trọng hoặc có rủi ro cao. Việc phát hành những đánh giá này là một bước tích cực hướng tới xây dựng niềm tin và sự tự tin vào các thực hành an ninh mạng của OpenAI.

TIPS & TRICKS CHO DEV

Tối ưu hóa code

Vấn đề: Code `messy` và khó đọc.

Cách làm: Sử dụng Claude Code với lệnh `claude refactor` để tối ưu hóa code. Ví dụ: `claude refactor --language python --input-file example.py`.

Đánh giá: Hiệu quả cao, nên dùng khi code cần tối ưu hóa.

Debug nhanh chóng

Vấn đề: Debug code tốn nhiều thời gian.

Cách làm: Sử dụng Claude Code với lệnh `claude debug` để debug nhanh chóng. Ví dụ: `claude debug --language python --input-file example.py`.

Đánh giá: Nhanh chóng và hiệu quả, nên dùng khi gặp lỗi.

Tạo code nhanh

Vấn đề: Tạo code từ scratch tốn nhiều thời gian.

Cách làm: Sử dụng Claude Code với lệnh `claude generate` để tạo code nhanh chóng. Ví dụ: `claude generate --language python --prompt "tạo hàm tính"`

tổng".

Đánh giá: Hiệu quả cao, nên dùng khi cần tạo code nhanh.

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 mô hình AI. Điều này giúp cải thiện hiệu suất của ứng dụng và giảm thiểu gánh nặng tài nguyên.

2. Việc tối ưu hóa LLM có thể được thực hiện thông qua việc fine-tuning và sử dụng các kỹ thuật như LoRA (Low-Rank Adaptation) để giảm kích thước mô hình và tăng tốc độ xử lý.

3. Ví dụ, khi sử dụng thư viện Hugging Face Transformers, dev có thể áp dụng kỹ thuật LoRA để tối ưu hóa mô hình LLM như sau:

```
from transformers import AutoModelForCausalLM, AutoTokenizer
```

Tải mô hình và tokenizer

```
model = AutoModelForCausalLM.from_pretrained("model-name")
```

```
tokenizer = AutoTokenizer.from_pretrained("model-name")
```

Áp dụng LoRA để tối ưu hóa mô hình

```
model = model.apply_lora()
```

4. Tip: Để tối ưu hóa hiệu năng LLM, dev nên thử nghiệm với các kỹ thuật khác nhau như quantization, pruning và knowledge distillation để tìm ra phương pháp phù hợp nhất với nhu cầu của ứng dụng.

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

Nguồn: Google News · Groq AI