

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Khó mà không khổ, khổ mà không khó.”

— Tục ngữ Việt Nam

Nghèo khó về vật chất không đồng nghĩa với khổ sở tinh thần — thái độ sống tích cực mới quyết định hạnh phúc.

TIN TỨC NỔI BẬT

1

Xây dựng agent phân tích tài chính thông minh với LangGraph và Strands Agents

Build an intelligent financial analysis agent with LangGraph and Strands Agents

Amazon Web Services (AWS)

[Đọc bài viết →](#)

Amazon Web Services (AWS) đã giới thiệu một khả năng mới để xây dựng một agent phân tích tài chính thông minh sử dụng LangGraph và Strands Agents. Giải pháp sáng tạo này cho phép người dùng tạo ra một công cụ phân tích tài chính tinh vi có thể xử lý và phân tích lượng lớn dữ liệu tài chính. LangGraph, một thư viện xử lý ngôn ngữ tự nhiên (NLP), được sử dụng để trích xuất thông tin liên quan từ các tài liệu tài chính, chẳng hạn như báo cáo và bảng kê. Strands Agents, một nền tảng low-code, được sử dụng để xây dựng và triển khai agent phân tích tài chính. Với agent này, người dùng có thể tự động hóa các tác vụ phân tích tài chính, như trích xuất, xử lý và trực quan hóa dữ liệu. Agent cũng có thể cung cấp các insight và khuyến nghị dựa trên dữ liệu đã phân tích, giúp người dùng đưa ra các quyết định tài chính sáng suốt. Agent phân tích tài chính thông minh được xây dựng trên AWS, cung cấp khả năng mở rộng, bảo mật và độ tin cậy. Giải pháp này đặc biệt hữu ích cho các tổ chức tài chính, công ty kế toán và các doanh nghiệp yêu cầu khả năng phân tích tài chính nâng cao. Bằng cách tận dụng LangGraph và Strands Agents, người dùng có thể tinh gọn quy trình phân tích tài chính của mình và thu được những insight giá trị về hiệu suất tài chính.

2

LangChain ra mắt công cụ xây dựng agent mã nguồn mở "dễ tiếp cận"

LangChain launches "accessible" open source agent builder

thestack.technology

[Đọc bài viết →](#)

LangChain đã giới thiệu một công cụ xây dựng agent mã nguồn mở, nhằm mục đích giúp các developer dễ dàng hơn trong việc tạo và tích hợp các agent được hỗ trợ bởi AI vào ứng dụng của họ. Công cụ xây dựng agent được thiết kế để đơn giản hóa quy trình xây dựng và triển khai các model AI đàm thoại, cho phép các developer tập trung vào việc tích hợp các model này vào sản phẩm của họ mà không cần chuyên môn sâu rộng về phát triển AI. Bản chất mã nguồn mở của công cụ xây dựng agent được kỳ vọng sẽ thúc đẩy sự hợp tác và phát triển dựa vào cộng đồng, cho phép người dùng đóng góp vào dự án và nâng cao khả năng của nó. Mục tiêu của LangChain là làm cho AI dễ tiếp cận hơn với nhiều developer, bất kể nền tảng hay kinh nghiệm của họ trong phát triển AI. Công cụ xây dựng agent được thiết kế để hoạt động liền mạch với nhiều ngôn ngữ lập trình và framework khác nhau, giúp các developer dễ dàng tích hợp các agent được hỗ trợ bởi AI vào ứng dụng của họ. Bằng cách cung cấp một giải pháp dễ tiếp cận và thân thiện với người dùng hơn, công cụ xây dựng agent của LangChain có tiềm năng đẩy nhanh việc áp dụng AI trong nhiều ngành và ứng dụng khác nhau.

3

Các model "open-weight" tốt nhất của Trung Quốc — và những đối thủ mạnh nhất từ Mỹ

The best Chinese open-weight models — and the strongest US rivals

understandingai.org

[Đọc bài viết →](#)

Bài viết so sánh các model "open-weight" của Trung Quốc với các đối thủ mạnh nhất từ Mỹ. Các model "open-weight" ở đây đề cập đến các robot được thiết kế để nâng và di chuyển các vật nặng, thường được sử dụng trong môi trường công nghiệp. Các nhà sản xuất Trung Quốc đã đạt được những bước tiến đáng kể trong lĩnh vực này, sản xuất các model chất lượng cao và hiệu quả về chi phí. Một số model "open-weight" hàng đầu của Trung Quốc bao gồm FANUC CRX, KUKA KR CYBERTEC và Yaskawa Motoman. Các robot này nổi tiếng về độ chính xác, tốc độ và độ tin cậy. Ngược lại, các công ty có trụ sở tại Mỹ như ABB, KUKA và FANUC cũng cung cấp các model "open-weight" hiệu suất cao. IRB 6700 của ABB và LBR iiwa của KUKA là những ví dụ đáng chú ý, cung cấp các tính năng và khả năng tiên tiến. Bài viết nhấn mạnh sự cạnh tranh ngày càng tăng giữa các nhà sản xuất Trung

Quốc và Mỹ trên thị trường robot "open-weight". Các công ty Trung Quốc đang dần chiếm ưu thế nhờ giá cả cạnh tranh và thiết kế sáng tạo, trong khi các công ty Mỹ vẫn duy trì thị phần thông qua danh tiếng đã được thiết lập và công nghệ tiên tiến.

4

"Phim" AI zombie mới của Neill Blomkamp chỉ là món cũ xào lại

Neill Blomkamp's new zombie AI 'film' is just slop warmed over

The Verge AI [Đọc bài viết →](#)

Neill Blomkamp, đạo diễn của District 9 và Gran Turismo, đã công bố dự án mới nhất của mình, một phim ngắn khoa học viễn tưởng dài 13 phút mang tên Nightborne. Phim ngắn này dựa trên tiểu thuyết Echopraxia năm 2014 của Peter Watts và có các nhân vật với giọng nói và khuôn mặt được mô phỏng theo diễn viên người thật. Tuy nhiên, mọi cảnh quay trong Nightborne đều được tạo ra bằng trình tạo video từ văn bản Seedance 2.0 của ByteDance. Blomkamp coi Nightborne là một "khởi đầu thử nghiệm" để chứng minh khả năng của generative AI và đặt mục tiêu tạo ra một bộ phim truyện đầy đủ theo định dạng này trong tương lai. Mặc dù được trau chuốt hơn so với các video do AI tạo ra khác, Nightborne vẫn thể hiện rõ những đặc điểm của nội dung do máy móc tạo ra. Bộ phim kể về một cuộc xung đột quân sự và một chương trình của chính phủ biến những người lính chết não thành zombie điều khiển từ xa. Mặc dù một nhóm người thật đã tham gia vào quá trình sản xuất dự án, Nightborne được mô tả là một "trải nghiệm xem tệ hại" và sự phụ thuộc của nó vào nội dung do AI tạo ra thể hiện rõ trong hình ảnh và âm thanh.

5

TreeSize sẽ không gia hạn hỗ trợ giấy phép vĩnh viễn trừ khi người dùng đăng ký gói thuê bao

TreeSize won't renew perpetual-license support unless users subscribe

Ars Technica [Đọc bài viết →](#)

TreeSize, một công cụ phân tích dung lượng ổ đĩa, đã thay đổi mô hình kinh doanh của mình để đáp ứng "điều kiện kinh tế hiện tại". Công ty, JAM Software, hiện yêu cầu người dùng có giấy phép vĩnh viễn phải đăng ký gói thuê bao để nhận hỗ trợ và cập nhật sau thời gian bảo trì 12 tháng ban đầu. Thay đổi này đã gây ra sự thất vọng cho người dùng, những người được yêu cầu sao lưu các file cài đặt và license key

của họ để tránh mất quyền truy cập vào phần mềm. JAM Software tuyên bố chính sách này đã có hiệu lực từ năm 2025, nhưng người dùng hiện đang phát hiện ra nó khi thời gian bảo trì của họ hết hạn. Công ty lập luận rằng việc cung cấp các file cài đặt sao lưu và license key tiềm ẩn rủi ro do các lỗ hổng bảo mật và chi phí bảo trì. Giám đốc sản phẩm của TreeSize, Hendrik Christ, tuyên bố rằng sự thay đổi là cần thiết để đảm bảo công ty có thể tiếp tục phát triển và hỗ trợ phần mềm về lâu dài. Tuy nhiên, một số người dùng đang đặt câu hỏi về sự cần thiết phải đăng ký gói thuê bao cho một công cụ phân tích dung lượng ổ đĩa, đặc biệt khi có các lựa chọn thay thế miễn phí.

6

Tại sao AI cần một "Hệ số Genie"

Why AI Needs a "Genie Coefficient"

IEEE Spectrum

[Đọc bài viết →](#)

Một nhóm chuyên gia, bao gồm Barath Raghavan và Bruce Schneier, đề xuất một metric mới gọi là "Hệ số Genie" để đo lường liệu các hệ thống AI có thực sự làm những gì người dùng mong muốn hay không. Các benchmark hiện tại tập trung vào những gì AI có thể làm, chứ không phải liệu nó có đạt được mục tiêu dự định của người dùng hay không. Hệ số Genie nhằm mục đích thu hẹp khoảng cách giữa yêu cầu của người dùng và sự hiểu biết của AI, vốn thường bị ảnh hưởng bởi kiến thức chung và ngữ cảnh. Điều này là do ngôn ngữ con người vốn dĩ không được xác định rõ ràng, khiến việc liệt kê tất cả các cảnh báo và ngoại lệ là không thể. Tuy nhiên, con người có thể đưa ra những phỏng đoán hợp lý dựa trên văn hóa chung, giao tiếp trước đó và hành vi bẩm sinh. Mặt khác, các AI agent có thể hiểu sai yêu cầu do thiếu hiểu biết về ngữ cảnh con người. Với việc sử dụng ngày càng nhiều các "harness" linh hoạt, các AI agent đang trở nên chủ động và tự chủ hơn, điều này có thể dẫn đến hành vi không mong muốn và có khả năng gây ra vấn đề. Các tác giả lập luận rằng cần có một metric mới để đảm bảo các hệ thống AI phù hợp với ý định của người dùng và tránh những hậu quả không mong muốn.

7

Apollo sử dụng Deep Agents và LangSmith cho AI GTM như thế nào

How Apollo Uses Deep Agents and LangSmith for GTM AI

LangChain Blog

[Đọc bài viết →](#)

Apollo, một nền tảng go-to-market (GTM), đã cải tiến AI Assistant của mình để cung cấp trải nghiệm người dùng hợp lý hơn. Nền tảng này, bao gồm toàn bộ chu trình bán hàng, đã trở nên quá phức tạp đối với khách hàng do phạm vi tính năng rộng lớn. Apollo đặt mục tiêu đơn giản hóa quy trình bằng cách giới thiệu một giao diện dựa trên chat, nơi người dùng có thể nêu mục tiêu bằng ngôn ngữ tự nhiên và trợ lý sẽ thực hiện các tác vụ cần thiết. Ban đầu, nền tảng sử dụng kiến trúc dựa trên supervisor với LangGraph, nhưng điều này tỏ ra hạn chế. Nhóm sau đó đã đánh giá hai lựa chọn thay thế, chọn Deep Agents vì tính trung lập về model của nó, cho phép Apollo sử dụng nhiều nhà cung cấp Large Language Model (LLM). Kiến trúc mới cho phép trợ lý tự động chọn các skill dựa trên mục tiêu của người dùng, tự tổng hợp kế hoạch và thực hiện các tác vụ mà không yêu cầu chỉ định trước. Điều này đã giúp cải thiện độ trễ, giảm các lời nhắc xác nhận và rút ngắn chu trình phát triển. Ngoài ra, Apollo đã bắt đầu sử dụng LangSmith để quan sát và debug, cho phép các kỹ sư theo dõi các vấn đề và hiểu rõ hoạt động bên trong của hệ thống AI.

Ba điểm yếu khiến bảo mật doanh nghiệp thất bại ở quy mô codebase (và tại sao các công cụ hiện tại của bạn không giải quyết được)

Three places enterprise security breaks down at codebase scale (and why your current tools don't cover them)

Sourcegraph Blog

[Đọc bài viết →](#)

Bảo mật doanh nghiệp thường bị xâm phạm ở quy mô codebase do một số lỗ hổng chính. Một vấn đề lớn là sự thất bại trong việc khắc phục lỗ hổng, vốn bị cản trở bởi phân tích code AI thiếu ngữ cảnh. Điều này gây khó khăn cho các nhóm bảo mật trong việc xác định và giải quyết chính xác các mối đe dọa tiềm tàng. Hơn nữa, tồn tại những khoảng trống trong phạm vi phát hiện, tạo ra các điểm mù trong bức tranh bảo mật. Những khoảng trống này có thể đặc biệt có vấn đề khi xử lý các codebase lớn, chẳng hạn như những codebase được tìm thấy trong 10.000 repository. Ngoài ra, thời gian để triển khai các bản sửa lỗi có thể kéo dài quá mức, thường mất hàng tuần để giải quyết các vấn đề. Những thách thức này làm nổi bật những hạn chế của các công cụ bảo mật hiện tại và nhu cầu về các giải pháp hiệu quả hơn có thể mở rộng để đáp ứng nhu cầu của các codebase lớn. Bằng cách giải quyết những lỗ hổng này, các tổ chức có thể củng cố tư thế bảo mật của mình và bảo vệ tốt hơn trước các mối đe dọa tiềm tàng.

TIPS & TRICKS CHO DEV

Tạo Agents Đa Nhiệm

Vấn đề: Xử lý task phức tạp cần nhiều AI agents.

Cách làm: Sử dụng LangGraph, CrewAI để tạo agents, ví dụ `langgraph create-agent --name my_agent`.

Đánh giá: Hiệu quả cao, nên dùng khi cần phân tích dữ liệu đa chiều.

Tích Hợp AutoGen

Vấn đề: Cần tự động hóa quá trình tạo agents.

Cách làm: Sử dụng AutoGen, ví dụ `autogen generate --template agent`.

Đánh giá: Tiết kiệm thời gian, nên dùng khi cần tạo nhiều agents tương tự.

Quản Lý Agents

Vấn đề: Khó quản lý và phối hợp nhiều AI agents.

Cách làm: Sử dụng phiData để quản lý và theo dõi agents, ví dụ `phidata list-agents`.

Đánh giá: Dễ dàng quản lý, nên dùng khi cần theo dõi hiệu suất của nhiều agents.

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 giảm thiểu tài nguyên và tăng tốc độ xử lý trong các ứng dụng AI. Điều này đặc biệt quan trọng khi triển khai mô hình AI trên các thiết bị edge hoặc trong các hệ thống thực thời.
3. Ví dụ, sử dụng kỹ thuật model pruning và quantization có thể giảm kích thước mô hình LLM và tăng tốc độ inference.
4. Tip: Sử dụng thư viện như Hugging Face Transformers để thực hiện fine-tuning và tối ưu hóa mô hình LLM cho các use case cụ thể, giúp giảm chi phí và tăng hiệu năng.

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

Nguồn: Google News · Groq AI