BIS - MT · 19/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

> *"Mưu sự tại nhân, thành sự tại thiên."*
>
> — Tục ngữ Việt Nam

Hãy hết lòng nỗ lực và lên kế hoạch chu đáo — còn kết quả cuối cùng tùy thuộc vào nhiều yếu tố khách quan.

TIN TỨC NỔI BẬT

1 Claude Code vs GitHub Copilot 2026: SWE-bench, Giá cả [Đã thử nghiệm]

Claude Code vs GitHub Copilot 2026: SWE-bench, Pricing [Tested]

tech-insider.org Đọc bài viết →

Trong một so sánh gần đây, Claude Code và GitHub Copilot đã được thử nghiệm trong một điểm chuẩn Kỹ thuật Phần mềm (SWE). Kết quả cho thấy Claude Code vượt trội so với GitHub Copilot ở một số lĩnh vực chính. Mặc dù cả hai công cụ mã hóa được hỗ trợ bởi AI đều nhằm giúp các nhà phát triển với việc hoàn thành mã và đề xuất, Claude Code đã thể hiện một cách tiếp cận chính xác và hiệu quả hơn. Bài kiểm tra SWE-bench, đánh giá một loạt các nhiệm vụ lập trình, đã tiết lộ rằng Claude Code hoàn thành nhiệm vụ nhanh hơn và với ít lỗi hơn. Về giá cả, cả hai công cụ đều cung cấp các mô hình khác nhau. GitHub Copilot yêu cầu đăng ký kế hoạch Nhà phát triển của GitHub, với giá 19,99 đô la mỗi tháng, trong khi Claude Code cung cấp một hệ thống giá nhiều tầng, với một kế hoạch miễn phí có sẵn cho sử dụng cá nhân và các tính năng tiên tiến hơn có sẵn với chi phí cao hơn. Các chi tiết giá chính xác cho Claude Code không được chỉ định trong so sánh. Tổng thể, kết quả kiểm tra cho thấy Claude Code có thể là một lựa chọn hiệu quả hơn cho các nhà phát triển tìm kiếm sự hỗ trợ mã hóa được hỗ trợ bởi AI.

2 Lớp lệnh AI Multi-Agent cho kho hàng giúp tối ưu vận hành và 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, được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng. Công nghệ đổi mới này kết hợp trí tuệ nhân tạo (AI) và học máy (ML) để tối ưu hóa hoạt động nhà kho, cho phép các doanh nghiệp tối ưu hóa hoạt động hậu cần và đưa ra quyết định dựa trên dữ liệu. Lớp Lệnh AI Nhà kho Đa tác nhân là một trung tâm điều khiển tích hợp nhiều tác nhân AI, mỗi tác nhân chuyên về các nhiệm vụ khác nhau như quản lý hàng tồn kho, thực hiện đơn hàng và tối ưu hóa chuỗi cung ứng. Những tác nhân này làm việc cùng nhau để phân tích dữ liệu từ các nguồn khác nhau, bao gồm cảm biến, camera và thiết bị IoT, để cung cấp thông tin và khuyến nghị theo thời gian thực. Bằng cách tận dụng công nghệ này, các doanh nghiệp có thể cải thiện hiệu quả hoạt động, giảm chi phí và nâng cao sự hài lòng của khách hàng. Lớp Lệnh AI Nhà kho Đa tác nhân cũng có thể giúp xác định các điểm nghẽn và khu vực cần cải thiện, cho phép các tổ chức đưa ra quyết định dựa trên dữ liệu và thúc đẩy tăng trưởng kinh doanh. Giải pháp này được xây dựng trên nền tảng AI của NVIDIA, cho phép tích hợp liền mạch với các hệ thống hiện có và khả năng mở rộng để đáp ứng nhu cầu của các doanh nghiệp lớn.

3 AWS open-source MCP server cho Bedrock AgentCore nhằm tinh gọn phát triển AI agent

AWS Open-Sources an MCP Server for Bedrock AgentCore to Streamline AI Agent Development

MarkTechPost Đọc bài viết →

Amazon Web Services (AWS) đã mở nguồn một máy chủ MCP (Nền tảng Đa đám mây) cho Bedrock AgentCore, một framework được thiết kế để đơn giản hóa việc phát triển đại lý AI. Động thái này nhằm mục đích đơn giản hóa quá trình tạo và triển khai các đại lý AI trên nhiều nền tảng đám mây khác nhau. Bedrock AgentCore là một framework không phụ thuộc vào đám mây, cho phép các nhà phát triển xây dựng, đào tạo và triển khai các đại lý AI trên các dịch vụ đám mây khác nhau. Máy chủ MCP mã nguồn mở được thiết kế để hoạt động liền mạch với Bedrock AgentCore, cho phép các nhà phát triển quản lý và điều phối các đại lý AI trên nhiều môi trường đám mây. Bằng cách mở nguồn máy chủ MCP, AWS đang cung cấp một nền tảng được thúc đẩy bởi cộng đồng cho các nhà phát triển đóng góp và cải thiện

framework. Động thái này dự kiến sẽ đẩy nhanh sự phát triển của các đại lý AI và thúc đẩy sự hợp tác giữa các nhà phát triển, nhà nghiên cứu và các tổ chức làm việc trên các dự án liên quan đến AI. Máy chủ MCP mã nguồn mở hiện đã có sẵn trên GitHub, cho phép các nhà phát triển truy cập và sử dụng framework cho nhu cầu phát triển đại lý AI của họ.

4

Nữ nghiên cứu sinh trang bị kỹ năng lắp ráp cho robot của NASA

This Graduate Student Equips NASA's Robots With Assembly Skills

IEEE Spectrum Đọc bài viết →

Sarah Downs, một sinh viên sau đại học về kỹ thuật điện, đã phát triển một thuật toán hợp tác với NASA và Không quân Hoa Kỳ cho phép robot lắp ráp vệ tinh trong không gian. Thuật toán này giải quyết vấn đề kinh điển "peg-in-hole" bằng cách cho phép robot chèn một ăng-ten vào vị trí chính xác trên vệ tinh. Sự quan tâm của Downs đến lĩnh vực robot bắt đầu từ khi còn nhỏ, được kích thích bởi việc tham gia đội First Lego League ở trường trung học và cảm hứng từ các nhiệm vụ rover sao Hỏa của NASA. Cô theo đuổi sự nghiệp STEM, kiếm được bằng cử nhân về kỹ thuật điện từ Đại học Tulsa và tiếp tục học tập như một sinh viên tiến sĩ tại Đại học Texas A&M. Nghiên cứu của Downs tập trung vào việc lắp ráp và điều khiển vệ tinh, với mục tiêu phát triển robot có thể hỗ trợ trong các nhiệm vụ không gian. Công việc của cô với NASA là kết quả của giấc mơ làm việc với cơ quan vũ trụ, điều mà cô đã khát khao từ những năm tuổi thiếu niên.

5

Từ các dự án open source thành công đến OpenAI

From open source hits to OpenAI

Changelog Đọc bài viết →

Trong tập này, Max Stoiber, một developer làm việc trên thư mục plugin và nền tảng ứng dụng của ChatGPT tại OpenAI, chia sẻ những nhận xét của mình về ngành công nghệ. Ông thảo luận về các dự án mã nguồn mở thường bị bỏ qua, chẳng hạn như react-boilerplate và styled-components, đã có tác động đáng kể. Stoiber cũng nói về sự phát triển của Spectrum, một nền tảng đã trở thành một phần của GitHub và đóng góp vào sự phát triển của GitHub Discussions. Ngoài ra, ông chia sẻ kinh nghiệm xây dựng Stellate, một bộ nhớ đệm

GraphQL đã được Shopify và The Guild mua lại. Stoiber tin rằng các ứng dụng ChatGPT đại diện cho một bề mặt mới cho phát triển phần mềm. Tập này có sẵn cho các thành viên Changelog++, những người có thể truy cập mà không có quảng cáo.

6

Dave Eggers nói với nhân viên OpenAI rằng ChatGPT đang 'làm câm lặng cả một thế hệ'

Dave Eggers told OpenAI staff that ChatGPT was ‘silencing an entire generation’

The Verge AI Đọc bài viết →

Tác giả Dave Eggers gần đây đã có một bài nói chuyện với khoảng 200 nhân viên OpenAI, được mời bởi Sam Altman. Thay vì đưa ra lời khuyên về năng suất hoặc vượt trội trong nhiều lĩnh vực, Eggers đã chỉ trích công ty. Theo Financial Times, Eggers cho rằng ChatGPT đã làm cho cuộc sống của các nhà giáo dục trở nên "tai họa" bằng cách giúp học sinh dễ dàng sử dụng nội dung được tạo ra bởi AI, từ đó ngăn cản họ học cách viết và thể hiện bản thân. Eggers tin rằng điều này sẽ "làm im lặng một hoặc hai thế hệ." Việc chỉ trích này không có gì ngạc nhiên, xét đến công việc trong quá khứ của Eggers, bao gồm cả tiểu thuyết "The Circle", một lời chỉ trích gay gắt đối với ngành công nghệ, và mô tả trước đó của ông về việc viết được tạo ra bởi AI là "đồ giả tạo vô nghĩa".

7

Fine-tune các video và image model quy mô lớn với NVIDIA NeMo Automodel và Diffusers

Fine-tune video and image models at scale with NVIDIA NeMo Automodel and Diffusers

Hugging Face Blog Đọc bài viết →

NVIDIA và Hugging Face đã hợp tác để mang lại đào tạo phân tán cấp sản xuất cho bất kỳ mô hình nào ở định dạng Diffusers trên Hugging Face Hub. Tích hợp này, được hỗ trợ bởi thư viện mã nguồn mở NVIDIA NeMo Automodel, cho phép tinh chỉnh mô hình video và hình ảnh với quy mô lớn mà không cần chuyển đổi điểm kiểm tra hoặc viết lại mô hình. NeMo Automodel là một thư viện đào tạo bản địa PyTorch DTensor hỗ trợ mô hình khớp dòng và sử dụng khớp dòng làm mục tiêu đào tạo, với đào tạo không gian và đa phân giải bucketed dataloading để tăng tốc độ truyền tải. Tích hợp này cung cấp một số lợi ích, bao gồm không cần chuyển đổi điểm kiểm tra, trọng số được đào tạo trước hoạt động ngay lập tức và đường dẫn nhanh đến hỗ trợ

mô hình mới. Nó cũng hỗ trợ tinh chỉnh đầy đủ và tinh chỉnh hiệu quả về tham số, cũng như đào tạo có thể mở rộng với các lược đồ phân mảnh và điều phối đa nút. Lưu trình điển hình để tinh chỉnh mô hình bao gồm mã hóa trước dữ liệu tập, khởi chạy đào tạo với YAML FLUX hiện có, tạo ra từ điểm kiểm tra tinh chỉnh và đánh giá hiệu suất. Sự hợp tác này mở khóa tiềm năng cho các nhà nghiên cứu và nhà xây dựng để tinh chỉnh mô hình video và hình ảnh với quy mô lớn, với khả năng đào tạo mô hình lớn hơn như FLUX.1-dev (12B) và HunyuanVideo (13B). Tích hợp này hoàn toàn mã nguồn mở theo Apache 2.0 và được tài liệu trong hướng dẫn đào tạo Diffusers.

8

Ba điểm yếu khiến enterprise security thất bại ở quy mô codebase (và lý do các tool hiện tại 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 tổn thương ở quy mô codebase do một số vấn đề chính. Một lĩnh vực quan ngại lớn là khắc phục lỗ hổng, bị cản trở bởi việc sử dụng mã AI mà không có ngữ cảnh phù hợp. Sự thiếu hụt ngữ cảnh này có thể dẫn đến các bản vá không chính xác hoặc không đầy đủ, khiến việc giải quyết các mối đe dọa bảo mật trở nên khó khăn. Một thách thức khác là khoảng trống trong phạm vi phát hiện, nơi các công cụ bảo mật hiện tại không thể xác định các lỗ hổng trên một codebase lớn. Điều này có thể để lại các hệ thống doanh nghiệp dễ bị tấn công, làm cho việc giải quyết các khoảng trống này trở nên cần thiết. Cuối cùng, quá trình áp dụng các bản vá có thể tốn thời gian và đòi hỏi nhiều lao động, mất vài tuần để triển khai trên 10.000 kho lưu trữ. Quá trình khắc phục kéo dài này có thể để lại các hệ thống doanh nghiệp dễ bị vi phạm bảo mật, nhấn mạnh nhu cầu về các giải pháp bảo mật hiệu quả và hiệu suất hơn.

TIPS & TRICKS CHO DEV

Tối ưu hóa mã với GitHub Copilot

Vấn đề: Mã nguồn không tối ưu, cần cải thiện hiệu suất.

Cách làm: Sử dụng GitHub Copilot để đề xuất mã tối ưu, nhập lệnh `git copilot suggest` và cung cấp mã nguồn cần cải thiện.

Đánh giá: Hiệu quả cao, giúp giảm thời gian và tăng chất lượng mã.

Tự động hoàn thiện mã với Cursor

Vấn đề: Nhập mã nguồn thủ công tốn thời gian.

Cách làm: Sử dụng Cursor để tự động hoàn thiện mã, nhập lệnh `cursor complete` và bắt đầu nhập mã.

Đánh giá: Tiết kiệm thời gian, tăng hiệu suất lập trình.

Debug mã với Aider

Vấn đề: Khó tìm ra lỗi trong mã nguồn.

Cách làm: Sử dụng Aider để debug mã, nhập lệnh `aider debug` và cung cấp mã nguồn cần kiểm tra.

Đánh giá: Hiệu quả cao, giúp tìm và sửa lỗi nhanh chóng.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để tối ưu hóa chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM), các nhà phát triển cần biết cách tinh chỉnh và tối ưu hóa mô hình cho các trường hợp sử dụng cụ thể. Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý.
3. Ví dụ, việc áp dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp cải thiện hiệu suất của mô hình trên các nhiệm vụ cụ thể.
4. Tip hoặc bước tiếp theo: Hãy bắt đầu bằng việc đánh giá nhu cầu và mục tiêu của dự án, sau đó áp dụng các kỹ thuật tinh chỉnh và tối ưu hóa mô hình để đạt được hiệu suất và chi phí tối ưu.