BIS - MT · 09/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“We know what we are, but know not what we may be.”

↪ Chúng ta biết mình là ai, nhưng chưa biết mình có thể trở thành ai.

— William Shakespeare

Tiềm năng con người là vô hạn — đừng bao giờ tự giới hạn bởi những gì bạn đã làm được, hãy khám phá những gì bạn có thể đạt được.

TIN TỨC NỔI BẬT

1

Mọi người đều muốn tham gia vào việc mã hóa theo cảm hứng — và Google không khác với Stitch, sản phẩm tiếp theo sau Jules

Everyone's looking to get in on vibe coding — and Google is no different with Stitch, its follow-up to Jules

VentureBeat Đọc bài viết →

> Google đang bước vào lĩnh vực mã hóa vibe với Stitch, phát triển mới nhất của họ. Mã hóa vibe để cập đến việc viết mã code mang tính thẩm mỹ và dễ đọc, thường ưu tiên phong cách hơn tính năng nghiêm ngặt. Xu hướng này đã trở nên phổ biến trong thời gian gần đây, với nhiều công ty và nhà phát triển cố gắng tận dụng nó. Stitch là sự tiếp nối của Jules, một công cụ khác của Google nhằm cải thiện khả năng đọc mã code. Mặc dù chi tiết cụ thể về Stitch còn hạn chế, việc phát hành nó cho thấy cam kết của Google trong việc làm cho mã hóa trở nên dễ tiếp cận và trực quan hơn. Sự tham gia của gã khổng lồ công nghệ này vào thị trường mã hóa vibe cho thấy sự công nhận ngày càng tăng về tầm quan trọng của thẩm mỹ mã code trong quá trình phát triển. Khi nhu cầu về mã hóa vibe tiếp tục tăng, sự tham gia của Google có thể giúp hợp pháp hóa xu hướng này và khuyến khích nhiều nhà phát triển hơn ưu tiên khả năng đọc mã code và phong cách. Tuy nhiên, các tính năng và khả năng chính xác của Stitch vẫn chưa được biết, khiến các nhà phát triển phải chờ đợi thêm thông tin về công cụ mới này.

2

Claude Code so với Cursor 2026: 80,8% SWE-bench, 1M Context [Đã thử nghiệm]

Claude Code vs Cursor 2026: 80.8% SWE-bench, 1M Context [Tested]

tech-insider.org Đọc bài viết →

Trong một bài kiểm tra hiệu suất gần đây, Claude Code và Cursor 2026 đã được so sánh với nhau để xác định khả năng hiệu suất của chúng. Kết quả cho thấy Claude Code đã đạt được điểm số ấn tượng 80,8% trên SWE-bench, một điểm chuẩn được sử dụng rộng rãi để đánh giá các model ngôn ngữ. Đây là một thành tựu đáng chú ý, vì SWE-bench kiểm tra khả năng của một model để hiểu và tạo ra văn bản giống con người. Bài kiểm tra cũng liên quan đến 1M context, ám chỉ một tập dữ liệu văn bản lớn mà các model đã được đào tạo. Dữ liệu đào tạo rộng lớn này cho phép các model học hỏi từ một loạt các nguồn và cải thiện khả năng hiểu và tạo ngôn ngữ của chúng. Kết quả của bài kiểm tra hiệu suất này cung cấp thông tin chi tiết quý giá về hiệu suất của Claude Code và Cursor 2026, và có thể sẽ thu hút sự quan tâm của các nhà phát triển và nhà nghiên cứu làm việc với các model ngôn ngữ. Cần phải có thêm thử nghiệm và đánh giá để hiểu đầy đủ về điểm mạnh và điểm yếu của các model này, nhưng bài kiểm tra ban đầu này cho thấy Claude Code là một ứng cử viên mạnh trong lĩnh vực xử lý ngôn ngữ.

3

Tối ưu hóa các quy trình làm việc trên GitHub với AI tạo sinh sử dụng Amazon Bedrock và MCP | Amazon Web Services

Streamline GitHub workflows with generative AI using Amazon Bedrock and MCP | Amazon Web Services

Amazon Web Services (AWS) Đọc bài viết →

Amazon Web Services (AWS) đã giới thiệu một tích hợp mới cho phép người dùng tối ưu hóa các quy trình làm việc trên GitHub bằng cách sử dụng trí tuệ nhân tạo tạo sinh (generative AI). Tích hợp này kết hợp Amazon Bedrock và MCP (Nền tảng Canvas Mô hình) để cung cấp trải nghiệm liền mạch cho các nhà phát triển. Amazon Bedrock là một dịch vụ cho phép người dùng tạo, đào tạo và triển khai các mô hình AI, trong khi MCP là một nền tảng cho phép người dùng xây dựng, triển khai và quản lý các mô hình AI. Bằng cách tích hợp hai dịch vụ này, người dùng có thể tận dụng trí tuệ nhân tạo tạo sinh để tự động hóa các nhiệm vụ và cải thiện quy trình làm việc trên GitHub. Với tích hợp này, các nhà phát triển có thể sử dụng Amazon Bedrock và MCP để tạo mã, tự động hóa kiểm tra và tối ưu hóa quy trình phát triển của họ.

Điều này có thể giúp giảm thời gian và công sức cần thiết để hoàn thành các nhiệm vụ, cho phép các nhà phát triển tập trung vào công việc phức tạp và sáng tạo hơn. Tích hợp này cũng cung cấp một giải pháp có thể mở rộng và bảo mật để quản lý các mô hình AI và quy trình làm việc, khiến nó trở thành một lựa chọn hấp dẫn cho các tổ chức muốn áp dụng các công cụ phát triển được hỗ trợ bởi AI.

4

GitHub cho người mới bắt đầu: Câu trả lời cho một số câu hỏi phổ biến

GitHub for Beginners: Answers to some common questions

GitHub Blog Đọc bài viết →

GitHub đã phát hành một loạt tài nguyên thân thiện với người mới bắt đầu để giúp các nhà phát triển bắt đầu với nền tảng này. Những tài nguyên này bao gồm một loạt các chủ đề, bao gồm trí tuệ nhân tạo và học máy, tạo mã AI, và các phương pháp hay nhất để xây dựng phần mềm với quy mô lớn. Nền tảng này cũng cung cấp tài nguyên để giúp các nhà phát triển phát triển kỹ năng và sự nghiệp của họ, cũng như thông tin về cách các nhóm kỹ thuật và bảo mật của GitHub sử dụng nền tảng để trở nên năng suất và hợp tác hơn. Những tài nguyên này cũng đi sâu vào khả năng và lợi ích của tạo mã AI, cách bắt đầu xây dựng, vận chuyển và bảo trì phần mềm với GitHub, và cách sử dụng tạo tăng cường bằng thu hồi (RAG) để thu thập thêm thông tin. Ngoài ra, GitHub đã giải quyết các thách thức phổ biến trong DevSecOps và cung cấp hướng dẫn về cách bắt đầu giải quyết chúng với AI và tự động hóa. Trong một tập gần đây của "GitHub cho người mới bắt đầu", nền tảng này đã trả lời một số câu hỏi phổ biến mà các nhà phát triển thường gặp khi mới bắt đầu. Những câu hỏi này bao gồm thông tin về các khóa SSH, được sử dụng để xác thực danh tính của nhà phát triển khi đẩy và kéo mã trên GitHub. Khóa riêng tư lưu trên máy tính của nhà phát triển, trong khi khóa công khai được chia sẻ với nền tảng.

5

Trí tuệ nhân tạo Siri tại WWDC 2026

Siri AI at WWDC 2026

Simon Willison Đọc bài viết →

Tại WWDC 2026 gần đây, Apple đã công bố các tính năng mới cho Siri AI của mình. Mặc dù một số người có thể hoài nghi sau các thông báo của năm ngoái, nhưng các tính năng mới này dường như khả thi với

công nghệ hiện tại. Apple đang cấp phép một mô hình AI tùy chỉnh, được sinh từ Gemini, để chạy trên Private Cloud Compute của mình. Điều này sẽ cho phép Siri trích xuất thông tin từ màn hình của người dùng bằng cách sử dụng các LLM tầm nhìn, loại bỏ nhu cầu về mã tùy chỉnh từ các ứng dụng hiện có. Thư viện Core AI mới cũng được giới thiệu, cho phép các nhà phát triển tận dụng phần cứng của Apple để chạy các mô hình của riêng họ. Nó tích hợp với hệ sinh thái PyTorch của Meta thông qua Core AI PyTorch Extensions, một gói Python kết nối PyTorch và Core AI. Các nhà phát triển có thể cài đặt iOS 27 Developer Beta để truy cập các tính năng mới, nhưng truy cập vào Siri AI mới hiện đang bị hạn chế.

6

MCP trên chế độ mã hóa

MCP on Code Mode

Changelog Đọc bài viết →

Trong một tập gần đây của MCP trên Code Mode, Matt Carey từ Cloudflare thảo luận về khái niệm MCP (Model-Code-Protocol) và cách nó đã bị hiểu lầm bởi nhiều người. Carey giải thích cách chế độ Code Mode phía máy chủ cho phép một máy chủ MCP duy nhất lộ hơn 2.500 điểm cuối API của Cloudflare bằng cách sử dụng chỉ 1.000 token ngữ cảnh. Ông cũng nói về trình tải Worker động mà chạy mã được viết bởi model một cách an toàn trong một môi trường cách ly V8. Ngoài ra, Carey chia sẻ quy trình làm việc cá nhân của mình với Claude, một model, và thảo luận về vai trò của bộ nhớ trong tương lai của các tác nhân. Tập này cũng giới thiệu các công cụ và nền tảng khác nhau, bao gồm Coder, Tailscale, RWX và Fly.io, được sử dụng cho môi trường phát triển an toàn, truy cập dựa trên danh tính và tự động hóa CI/CD.

7

Các nhà nghiên cứu đã đào tạo một tác nhân tìm kiếm AI mã nguồn mở, Harness-1, vượt trội so với GPT-5.4 trong việc nhớ lại thông tin liên quan

Researchers trained an open source AI search agent, Harness-1, that outperforms GPT-5.4 on recalling relevant information

VentureBeat Đọc bài viết →

Các nhà nghiên cứu tại Đại học Illinois tại Urbana-Champaign, UC Berkeley và Chroma đã phát triển một đại lý tìm kiếm AI mã nguồn mở gọi là Harness-1. Đại lý này vượt trội so với GPT-5.4 trong việc nhớ lại thông tin liên quan, đạt điểm trung bình là 73% so với 70,9% của

> GPT-5.4. Harness-1 cũng vượt trội so với các đại lý tìm kiếm mã nguồn mở khác, bao gồm Tongyi DeepResearch 30B, với 11,4 điểm phần trăm. Hiệu suất của mô hình này được cho là nhờ khả năng offload các nhiệm vụ quản lý thường xuyên, chẳng hạn như duy trì bộ nhớ làm việc và xác minh yêu cầu, vào một môi trường phần mềm có cấu trúc. Harness-1 được đào tạo bằng một phương pháp mới tách biệt các lựa chọn ngữ nghĩa khỏi quản lý trạng thái cấu trúc. Mô hình được dạy để vận hành một giao diện có cấu trúc, thay vì dựa vào việc ghi nhớ lực bạo các trạng thái tìm kiếm. Phương pháp này đã dẫn đến sự giảm đáng kể dữ liệu đào tạo, với mô hình được đào tạo trên chỉ 4.400 mục duy nhất. Mô hình Harness-1 có sẵn theo giấy phép Apache 2.0, làm cho nó có sẵn miễn phí cho sử dụng thương mại. Các nhà nghiên cứu tin rằng phương pháp này có tiềm năng cách mạng hóa cách thiết kế các hệ thống AI, chuyển sự tập trung từ đào tạo các mô hình lớn hơn sang xây dựng các môi trường tốt hơn cho các mô hình để hoạt động trong đó.

8

Tải xuống: cách quả bóng World Cup sẽ bay và ứng dụng "siêu" của OpenAI

The Download: how the World Cup ball will fly and OpenAI's "super app"

MIT Tech Review Đọc bài viết →

> Giải vô địch bóng đá thế giới FIFA sắp diễn ra, và giải đấu năm nay dự kiến sẽ là giải đấu khó đoán nhất từ trước đến nay. Các nhà nghiên cứu đã phát hiện ra rằng quả bóng Adidas Trionda mới, được thiết kế cho những cú sút xa, có thể không di chuyển xa như những quả bóng World Cup trước đây. Tuy nhiên, sự thay đổi này có thể dẫn đến một đường bay dự đoán được hơn, điều mà các cầu thủ không luôn thích. Trong các tin tức công nghệ khác, OpenAI đang lên kế hoạch biến ChatGPT thành một "siêu ứng dụng" trước khi IPO. Phiên bản được làm mới này sẽ kết hợp các công cụ mã hóa và các tác nhân AI, đánh dấu một sự thay đổi trong trọng tâm của công ty từ các rô-bốt trò chuyện sang các tác nhân AI. Ngoài ra, Google đã đồng ý trả cho SpaceX 30 tỷ đô la cho sức mạnh tính toán AI, trong khi AI dự kiến sẽ làm cho cuộc sống hàng ngày trở nên đắt hơn do sự thèm khát tài nguyên không thể thỏa mãn. Châu Âu cũng đang đẩy nhanh việc rút khỏi Big Tech của Mỹ, với hàng chục động thái chuyển sang các nhà cung cấp thay thế. Đây chỉ là một vài câu chuyện được đề cập trong phiên bản hôm nay của The Download, nơi các nhà phát triển có thể tìm hiểu về các API, LLM, và framework mới nhất, cũng như cách tích hợp AI vào mô hình và token của họ.

TIPS & TRICKS CHO DEV

Cài Đặt GitHub Copilot

Vấn đề: Thiết lập công cụ hỗ trợ viết mã chưa tối ưu.

Cách làm: Cài đặt GitHub Copilot trong IDE, nhập lệnh `copilot install` và thiết lập cấu hình.

Đánh giá: Tối ưu hóa quá trình viết mã, giúp tăng năng suất.

Sử Dụng Cursor

Vấn đề: Tìm kiếm và thay thế mã chưa hiệu quả.

Cách làm: Sử dụng công cụ như GitHub Cursor, nhập lệnh `cursor search` và prompt tìm kiếm.

Đánh giá: Tiết kiệm thời gian tìm kiếm và thay thế mã.

Tích Hợp Aider

Vấn đề: Thiếu hỗ trợ viết mã trong dự án.

Cách làm: Tích hợp Aider vào dự án, nhập lệnh `aider init` và thiết lập cấu hình.
Đánh giá: Cải thiện chất lượng mã và giảm thời gian viết mã.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Lập trình viên cần biết cách tối ưu hóa chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM) để đảm bảo ứng dụng của họ hoạt động hiệu quả và tiết kiệm. Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý. Ngoài ra, tối ưu hóa LLM cũng giúp cải thiện trải nghiệm người dùng.
3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning để điều chỉnh mô hình LLM cho phù hợp với nhiệm vụ cụ thể, hoặc sử dụng LoRA (Low-Rank Adaptation) để giảm thiểu kích thước mô hình.
4. Tip: Để bắt đầu tối ưu hóa LLM, hãy bắt đầu bằng cách phân tích hiệu suất của mô hình hiện tại và xác định các điểm cần cải thiện, sau đó áp dụng các kỹ thuật tối ưu hóa như fine-tuning, LoRA hoặc sử dụng các thư viện như Hugging Face's Transformers.