BIS - MT · 18/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Where there is a will, there is a way.”

↪ Có chí ắt có đường.

— William Hazlitt

Khi ý chí đủ mạnh, con người có thể tìm ra hoặc tạo ra con đường — đây là lý do tại sao thái độ và quyết tâm quan trọng hơn điều kiện.

TIN TỨC NỔI BẬT

1 Cẩm nang toàn diện về các file bộ nhớ của AI Agent (CLAUDE.md, AGENTS.md và hơn thế nữa)

The Complete Guide to AI Agent Memory Files (CLAUDE.md, AGENTS.md, and Beyond)

HackerNoon Đọc bài viết →

Bài viết cung cấp một hướng dẫn chi tiết để hiểu về các tệp tin bộ nhớ của tác nhân AI, tập trung cụ thể vào CLAUDE.md và AGENTS.md. Những tệp tin này rất quan trọng để lưu trữ và truy xuất thông tin về các tác nhân AI, cho phép các nhà phát triển quản lý và tương tác với chúng một cách hiệu quả hơn. CLAUDE.md là một tệp tin siêu dữ liệu chứa thông tin thiết yếu về một tác nhân AI, bao gồm tên, mô tả và các chi tiết liên quan khác. Tệp tin này đóng vai trò là trung tâm để truy cập và cập nhật siêu dữ liệu của một tác nhân. AGENTS.md, mặt khác, là một tệp tin lưu trữ danh sách tất cả các tác nhân AI trong một hệ thống. Nó cho phép các nhà phát triển dễ dàng quản lý và theo dõi các tác nhân của họ, giúp họ giữ cho dự án của mình được tổ chức tốt hơn. Bài viết nhấn mạnh tầm quan trọng của những tệp tin này trong phát triển AI, nhấn mạnh vai trò của chúng trong việc tạo điều kiện cho giao tiếp giữa con người và các tác nhân AI. Bằng cách hiểu CLAUDE.md và AGENTS.md, các nhà phát triển có thể tạo ra các hệ thống AI hiệu quả và hiệu suất hơn, dễ dàng quản lý và bảo trì hơn.

2 Ai cũng muốn tham gia vào vibe coding — và Google cũng không ngoại lệ với Stitch, công cụ kế nhiệm của 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 tham gia vào xu hướng "lập trình cảm giác" với việc giới thiệu Stitch, một công cụ mới đi theo bước chân của người tiền nhiệm, Jules. Lập trình cảm giác đề cập đến sự quan tâm ngày càng tăng trong việc tạo ra các công cụ lập trình ưu tiên trải nghiệm người dùng và thẩm mỹ, thay vì chỉ tập trung vào chức năng. Stitch được thiết kế để cung cấp một trải nghiệm lập trình trực quan và hấp dẫn về mặt hình ảnh, cho phép các developer tập trung vào các khía cạnh sáng tạo của lập trình mà không bị cản trở bởi cú pháp phức tạp và tổ chức mã. Mặc dù chi tiết về Stitch còn khan hiếm, việc phát hành nó cho thấy Google cam kết làm cho lập trình trở nên dễ tiếp cận và thú vị hơn cho một loạt người dùng rộng lớn hơn. Khi ngành công nghệ tiếp tục phát triển, rõ ràng rằng trải nghiệm người dùng đang trở thành ưu tiên hàng đầu. Với Stitch, Google đang thực hiện một bước tiến để làm cho lập trình trở nên dễ tiếp cận và thân thiện với người dùng hơn, có khả năng mở ra những cơ hội mới cho cả developer và những người không phải là developer.

3 Tổng quan về các Agent Framework: Chọn Runtime phù hợp cho việc triển khai AI trong doanh nghiệp

State of Agent Frameworks: Choosing the Right Runtime for Enterprise AI Execution

Medium Đọc bài viết →

> Bài viết "Trạng thái của các Framework Trình điều khiển: Chọn Thời gian chạy Phù hợp cho Thực thi AI Doanh nghiệp" cung cấp cái nhìn tổng quan về trạng thái hiện tại của các framework trình điều khiển trong bối cảnh thực thi AI doanh nghiệp. Các framework trình điều khiển là các nền tảng phần mềm cho phép phát triển và triển khai các trình điều khiển thông minh, là các chương trình phần mềm tự động có thể tương tác với môi trường và đưa ra quyết định. Bài viết nhấn mạnh tầm quan trọng của việc chọn thời gian chạy phù hợp cho thực thi AI doanh nghiệp, vì nó có thể ảnh hưởng đáng kể đến hiệu suất, khả năng mở rộng và khả năng bảo trì của các ứng dụng AI. Nó thảo luận về các framework trình điều khiển khác nhau, bao gồm Rasa, Botpress và Microsoft Bot Framework, cũng như các điểm mạnh và điểm yếu tương ứng của chúng. Bài viết cũng đề cập đến các yếu tố quan trọng khi chọn một framework trình điều khiển, chẳng hạn như dễ sử dụng, tùy chọn tùy chỉnh và tích hợp với các hệ thống hiện có. Nó nhấn mạnh nhu cầu về một framework có thể xử lý các cuộc trò

chuyện phức tạp, cung cấp bảo mật mạnh mẽ và hỗ trợ triển khai quy mô lớn. Cuối cùng, bài viết nhằm giúp các nhà phát triển và doanh nghiệp đưa ra quyết định thông minh khi chọn một framework trình điều khiển cho các ứng dụng AI của họ.

4

Tận dụng tối đa từng token: Cách Copilot cải thiện việc xử lý context và định tuyến model

Getting more from each token: How Copilot improves context handling and model routing

GitHub Blog Đọc bài viết →

GitHub Copilot đã cải thiện khả năng xử lý ngữ cảnh và định tuyến model, cho phép các developer thu được nhiều giá trị hơn từ mỗi token được sử dụng. Sự cải thiện này cho phép Copilot tập trung vào nhiệm vụ tại hand, thay vì lặp lại thông tin không cần thiết từ lượt này sang lượt khác. Hiệu suất tăng cường này có nghĩa là các developer có thể đạt được nhiều hơn với tín dụng của họ, vì Copilot thông minh hơn về cách sử dụng token. Để đạt được điều này, GitHub đang làm việc trên hai lĩnh vực chính: cải thiện bộ khung Copilot để chỉ đạo nhiều hơn mỗi phiên tới nhiệm vụ, và mở rộng tính năng Auto để cho phép Copilot tự động chọn model phù hợp nhất cho công việc. Sự phát triển này nhằm mục đích làm cho GitHub Copilot trở nên năng suất và hiệu quả hơn cho các developer, đặc biệt là trong các phiên dài hơn.

5

The Download: Cái nhìn thực tế về geoengineering và khoa học về interoception

The Download: a reality check for geoengineering and the science of interoception

MIT Tech Review Đọc bài viết →

Trong phiên bản ngày hôm nay của The Download, các nhà nghiên cứu đang làm việc về kỹ thuật địa chất mặt trời, một phương pháp để chống lại sự nóng lên toàn cầu, đang phải đối mặt với những thách thức kỹ thuật thực tế. Mặc dù đã đi vượt qua các mô hình mô phỏng trên máy tính, việc triển khai sớm sẽ yêu cầu cơ sở hạ tầng mới đáng kể, thời gian và đầu tư. Trong khi đó, các nhà khoa học đang đạt được tiến bộ trong việc hiểu về sự cảm nhận nội bộ, khả năng cảm nhận tín hiệu cơ thể nội bộ, nhờ một giải thưởng Nobel năm 2021 và các công cụ mới. Nghiên cứu này có ý nghĩa đối với việc điều trị các tình trạng như béo phì và lo lắng. Ngoài ra, bản tin bao gồm các câu chuyện công nghệ khác, bao gồm việc định giá tăng của SpaceX, các nhà lãnh

đạo G7 tìm kiếm quyền truy cập vào các mô hình AI hàng đầu của Mỹ, và một thiết bị cấy ghép não cho phép một bệnh nhân ALS không nói được làm việc toàn thời gian. Các chủ đề khác bao gồm sự trỗi dậy của các công ty khởi nghiệp chỉnh sửa gene, một vụ rò rỉ tiết lộ chi tiết về xã hội bí mật của Peter Thiel, và những tiến bộ trong tính toán lượng tử và AI.

6

AGI không phải là Multimodal

AGI Is Not Multimodal

The Gradient Đọc bài viết →

Những tiến bộ gần đây trong các mô hình AI tạo sinh đã khiến một số người tin rằng Trí tuệ Nhân tạo Tổng quát (AGI) đang sắp xảy ra. Tuy nhiên, những mô hình này đã xuất hiện không phải do các giải pháp sâu sắc, mà vì chúng đã được mở rộng hiệu quả trên phần cứng hiện có. Phương pháp đa mô thức, bao gồm việc tối ưu hóa các mạng mô-đun lớn cho nhiều mô thức, được coi là con đường dẫn đến AGI. Tuy nhiên, chiến lược này không thể thành công trong ngắn hạn và có thể không dẫn đến AGI ở mức độ con người. Một AGI thực sự phải là tổng quát trên tất cả các lĩnh vực và bao gồm khả năng giải quyết các vấn đề xuất phát từ thực tế vật lý, chẳng hạn như sửa chữa ô tô hoặc chuẩn bị thức ăn. Điều này đòi hỏi một hình thức trí tuệ cơ bản nằm trong một mô hình thế giới vật lý. Các Mô hình Ngôn ngữ Lớn (LLM) hiện tại đã được đề xuất để học một mô hình của thế giới thông qua dự đoán token tiếp theo, nhưng điều này có nhiều khả năng là một sự hiểu biết bề mặt về thực tế. Các khả năng của LLM đã dẫn đến sự nhầm lẫn về ý nghĩa của việc hiểu ngôn ngữ và thế giới.

7

Native Popover API: 4 Menu và Tooltip tôi xây dựng mà không cần JavaScript

The Native Popover API: 4 Menus and Tooltips I Built Without JavaScript

Dev.to AI Đọc bài viết →

Một nhà phát triển đã xây dựng thành công bốn thành phần UI cơ bản - dropdown, tooltip, menu lệnh và popup xác nhận - mà không dựa vào JavaScript. Thay vào đó, họ đã tận dụng API Popover gốc và định vị neo CSS để đạt được chức năng mong muốn. API Popover cung cấp khả năng kết xuất lớp trên cùng và khả năng hủy bỏ nhẹ, loại bỏ nhu cầu về các trình nghe sự kiện, biến trạng thái và thư viện. Cách tiếp

cận này đã giảm đáng kể độ phức tạp của mã và cải thiện hiệu suất. Nhà phát triển đã xây dựng bốn thành phần bằng cách sử dụng API Popover: một dropdown, một tooltip, một menu lệnh và một popup xác nhận. Dropdown và tooltip đã được tối ưu hóa đặc biệt, với logic định vị của dropdown được thay thế bằng một quy tắc CSS đơn giản và mã JavaScript của tooltip giảm 75%. API Popover cũng cho phép kết xuất lớp trên cùng, loại bỏ nhu cầu về các hack z-index và thư viện định vị. Cách tiếp cận mới này đã cải thiện khả năng truy cập và giảm công sức bảo trì. Trình duyệt xử lý toán học định vị, giảm các trường hợp biên và làm cho mã trở nên mạnh mẽ hơn. Nhà phát triển cũng lưu ý rằng hỗ trợ trình duyệt cho API Popover rất tốt, với Chrome, Edge, Safari và Firefox đều hỗ trợ nó.

8 Vụ rò rỉ dữ liệu lớn làm lộ credential của hàng ngàn mạng lưới nhạy cảm

Massive breach spills credentials for thousands of sensitive networks

Ars Technica Đọc bài viết →

Một sự vi phạm dữ liệu đáng kể đã được báo cáo, làm tổn thương các mạng nhạy cảm cho hàng nghìn tổ chức. Các bên bị ảnh hưởng bao gồm các công ty công nghệ lớn và một nhà thầu NATO. Oracle, một nhà cung cấp phần mềm doanh nghiệp hàng đầu, đã bị ảnh hưởng bởi sự vi phạm này. Lenovo, một nhà sản xuất máy tính cá nhân và laptop nổi tiếng, cũng đã bị ảnh hưởng. Sự vi phạm cũng mở rộng đến FedEx, một công ty hậu cần và giao hàng nổi bật. Ngoài các tập đoàn lớn này, một nhà thầu NATO cũng đã bị xâm phạm, làm dấy lên lo ngại về an ninh quốc gia. Fortinet, một công ty an ninh mạng, cũng nằm trong số các bên bị ảnh hưởng. Sự vi phạm đã□ lộ thông tin nhạy cảm, bao gồm thông tin đăng nhập mạng, có thể được sử dụng tiềm năng cho các hoạt động độc hại. Phạm vi đầy đủ của sự vi phạm và số lượng mạng bị xâm phạm vẫn chưa rõ ràng. Tuy nhiên, sự tham gia của các tổ chức có uy tín như vậy cho thấy một sự cố an ninh đáng kể và có khả năng lan rộng.

TIPS & TRICKS CHO DEV

Optimize Context Window

Vấn đề: Quản lý context window hiệu quả để tránh thông tin thừa.

Cách làm: Sử dụng kỹ thuật chunking, chia dữ liệu thành khối nhỏ hơn. Ví dụ, prompt "Tóm tắt văn bản dưới 100 từ" giúp hạn chế context window.

Đánh giá: Hiệu quả cao khi xử lý dữ liệu lớn, nhưng có thể mất thông tin quan trọng nếu không được thực hiện đúng.

Implement Long-Context

Vấn đề: Xử lý thông tin dài vượt quá giới hạn context window.
Cách làm: Sử dụng mô hình biến đổi attention mechanism, cho phép xử lý thông tin dài hơn. Ví dụ, lệnh CLI "transformers --model-type longformer" hỗ trợ long-context.
Đánh giá: Phù hợp cho ứng dụng cần xử lý văn bản dài, nhưng có thể tăng tải tính toán.

Manage Memory

Vấn đề: Quản lý bộ nhớ hiệu quả khi xử lý dữ liệu lớn.
Cách làm: Sử dụng kỹ thuật caching, lưu trữ thông tin trung gian để tránh tính toán lại. Ví dụ, prompt "Lưu trữ kết quả trung gian" giúp giảm tải bộ nhớ.
Đánh giá: Hiệu quả khi xử lý dữ liệu lớn, giảm thời gian phản hồi nhưng có thể tăng rủi ro mất dữ liệu nếu không được quản lý đúng.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để phát triển ứng dụng AI hiệu quả, dev cần biết cách tối ưu chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM). Việc này giúp giảm thiểu chi phí vận hành và tăng tốc độ xử lý dữ liệu. Điều này đặc biệt quan trọng khi triển khai ứng dụng trên các thiết bị di động hoặc edge AI.
3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và tăng tốc độ xử lý. Ví dụ code: `model = LLM.from_pretrained('base_model'); model.fine_tune(...)`
4. Tip: Để tối ưu hiệu năng LLM, hãy thử sử dụng các kỹ thuật như quantization, pruning và knowledge distillation. Ngoài ra, hãy xem xét việc sử dụng các framework như Hugging Face Transformers để đơn giản hóa quá trình tối ưu hóa.