BIS - MT · 09/07/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Đường dài mới biết ngựa hay, ở lâu mới biết người ngay kẻ tà.”

— Tục ngữ Việt Nam

Phẩm chất thật sự chỉ được bộc lộ qua thời gian và thử thách — đừng vội đánh giá người hay việc.

TIN TỨC NỔI BẬT

1

Hiện trạng các Agent Framework: Chọn Runtime phù hợp cho 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 tích hợp các model AI với các ứng dụng và hệ thống khác nhau. 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 hệ thống được hỗ trợ bởi AI. Nó thảo luận về các đặc điểm chính của các framework trình điều khiển, bao gồm khả năng xử lý các quy trình làm việc phức tạp, tích hợp với nhiều nguồn dữ liệu và cung cấp khả năng ra quyết định thời gian thực. Bài viết cũng khám phá các loại framework trình điều khiển khác nhau có sẵn, bao gồm framework dựa trên quy tắc, framework dựa trên học máy và framework hỗn hợp. Nó nhấn mạnh nhu cầu về một framework có thể thích nghi với các yêu cầu kinh doanh thay đổi và cung cấp mức độ tùy chỉnh và linh hoạt cao. Cuối cùng, bài viết nhằm giúp các tổ chức doanh nghiệp đưa ra quyết định thông minh khi chọn một framework trình điều khiển cho nhu cầu thực thi AI của họ, đảm bảo rằng họ chọn một nền tảng đáp ứng các yêu cầu cụ thể của họ và hỗ trợ các mục tiêu kinh doanh của họ.

2 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 giải pháp để xây dựng một tác nhân phân tích tài chính thông minh sử dụng LangGraph và Strands Agents. Cách tiếp cận đổi mới này cho phép người dùng tạo một công cụ phân tích tài chính tinh vi có thể xử lý và phân tích dữ liệu tài chính phức tạp. LangGraph là một model ngôn ngữ dựa trên đồ thị có thể hiểu và tạo ra văn bản giống con người. Nó được tích hợp với Strands Agents, một framework để xây dựng giao diện trò chuyện. Bằng cách kết hợp hai công nghệ này, người dùng có thể phát triển một tác nhân phân tích tài chính có thể cung cấp phân tích sâu sắc và chính xác về dữ liệu tài chính. Tác nhân này có thể được đào tạo trên các tập dữ liệu tài chính khác nhau, cho phép nó học các mẫu và mối quan hệ trong dữ liệu. Sau đó, nó có thể tạo báo cáo và cung cấp khuyến nghị cho người dùng dựa trên phân tích của nó. Giao diện trò chuyện của tác nhân cho phép người dùng tương tác với nó một cách tự nhiên và trực quan, giúp họ dễ dàng truy cập và hiểu rõ về thông tin tài chính. Giải pháp này có tiềm năng cách mạng hóa cách dữ liệu tài chính được phân tích và trình bày, cung cấp cho người dùng một sự hiểu biết toàn diện và có thể hành động hơn về tình hình tài chính của họ.

3 So sánh 7 LLM/hệ thống hàng đầu cho lập trình năm 2025

Comparing the Top 7 Large Language Models LLMs/Systems for Coding in 2025

MarkTechPost Đọc bài viết →

Bài viết so sánh 7 mô hình ngôn ngữ lớn (LLM) hàng đầu cho việc lập trình vào năm 2025. Những mô hình này đã cách mạng hóa lĩnh vực lập trình bằng cách cho phép các nhà phát triển viết mã code hiệu quả và chính xác hơn. Các mô hình được so sánh bao gồm LLaMA, PaLM 2, Chinchilla, Llama 2, BLOOM, OPT-175B và MPT-PM. Mỗi mô hình có điểm mạnh và điểm yếu độc đáo, với một số mô hình vượt trội trong các lĩnh vực cụ thể như hoàn thiện code, gỡ lỗi và tạo code. Bài viết nhấn mạnh khả năng của từng mô hình, bao gồm khả năng hiểu và tạo code trong nhiều ngôn ngữ lập trình khác nhau. So sánh cũng tính đến tài nguyên tính toán, dữ liệu đào tạo và khả năng tinh chỉnh của các mô hình. Thông tin này rất quan trọng đối với các nhà phát triển để

chọn mô hình phù hợp nhất với nhu cầu lập trình cụ thể của họ. Bằng cách hiểu điểm mạnh và hạn chế của từng mô hình, các nhà phát triển có thể tận dụng sức mạnh của LLM để cải thiện năng suất và hiệu quả lập trình của mình.

4

MCP trên Code Mode

MCP on Code Mode

Changelog Đọc bài viết →

Matt Carey của Cloudflare thảo luận về Code Mode và mối quan hệ của nó với MCP (Model-Code-Protocol) trong một cuộc phỏng vấn gần đây. Carey tiết lộ rằng chế độ Code Mode phía máy chủ cho phép một máy chủ MCP duy nhất hiển thị khoảng 2.500 điểm cuối API của Cloudflare bằng cách sử dụng khoảng 1.000 token ngữ cảnh. Ông cũng chia sẻ quy trình làm việc của mình với Claude, một model, và thảo luận về vai trò của bộ nhớ trong tương lai của các tác nhân. Ngoài ra, Carey đề cập đến trình tải Worker động, chạy an toàn mã được viết bởi model trong một môi trường cách ly V8. Cuộc trò chuyện cũng đề cập đến quy trình làm việc và công cụ cá nhân của Carey, chẳng hạn như trình bao git Zaggy của ông, ngăn chặn việc đẩy lực lên các kho lưu trữ của mình. Tập này được tài trợ bởi các công ty khác nhau, bao gồm Coder.com, Tailscale, RWX và Fly.io.

5

Giải mã loop engineering: Tận dụng tối đa AI agent, tránh "loopmaxxing"

Demystifying loop engineering: Get more from AI agents, avoid loopmaxxing

BD Tech Talks Đọc bài viết →

Các nhà phát triển đang thay đổi cách tiếp cận xây dựng ứng dụng với các mô hình ngôn ngữ lớn (LLM) bằng cách tập trung vào thiết kế các vòng lặp kích hoạt các tác nhân này, thay vì kích hoạt trực tiếp. Xu hướng này, được gọi là kỹ thuật vòng lặp, đã đang thu được động lực kể từ năm 2022 với sự xuất hiện của các vòng lặp suy luận kiểu ReAct. Thực tiễn này đã phát triển thông qua các thí nghiệm và sản phẩm hóa khác nhau, bao gồm các dự án mã nguồn mở AutoGPT và vòng lặp Ralph, và hiện đang được các nền tảng phát triển lớn áp dụng. Kỹ thuật vòng lặp đại diện cho một sự thay đổi□□ chức năng, chuyển từ việc quản lý vi mô các tương tác cá nhân sang việc tạo ra các dây chuyền lắp ráp phần mềm rộng lớn hơn. Điều này cho phép các hệ

thống đánh giá các đầu ra trung gian, cập nhật trạng thái của chúng và tự xác định các bước tiếp theo một cách tự chủ. Tuy nhiên, các vòng lặp được thiết kế kém có thể dẫn đến các vấn đề như chi phí tăng vọt, giảm khả năng quan sát và việc theo đuổi mục tiêu không hiệu quả. Để xây dựng các vòng lặp hiệu quả, các nhà phát triển cần tập trung vào thiết kế các chương trình có cấu trúc hoặc tự động hóa đã lên lịch để cung cấp ngữ cảnh và hướng dẫn cho các LLM, đánh giá đầu ra kết quả so với các tiêu chí bên ngoài và xác định xem nhiệm vụ có yêu cầu một lần lặp lại hay không. Điều này liên quan đến việc sử dụng đệ quy, theo dõi trạng thái bền vững, các plugin bên ngoài và các rào cản hoạt động nghiêm ngặt để tạo ra các vòng lặp sẵn sàng sản xuất.

6

STEM cần lãnh đạo từ mọi thế hệ tham gia

STEM Needs Leaders From Every Generation at the Table

IEEE Spectrum Đọc bài viết →

Hội nghị Lãnh đạo Quốc tế IEEE sẽ tập trung vào tầm quan trọng của lãnh đạo hợp tác trong các lĩnh vực STEM. Hội nghị, dự kiến diễn ra vào ngày 3-4 tháng 10 tại Budapest, nhằm trang bị cho các chuyên gia những công cụ cho lãnh đạo hợp tác và giải quyết nhu cầu chia sẻ kiến thức giữa các thế hệ. Với tốc độ phát triển công nghệ nhanh chóng và những thách thức toàn cầu phức tạp, lãnh đạo không còn là một nỗ lực cá nhân, mà là một nỗ lực hợp tác thúc đẩy kết nối, tận dụng các quan điểm đa dạng và hướng tới các kết quả chung. Hội nghị sẽ tập hợp các chuyên gia mới nổi, chuyên gia hàng đầu và các nhà lãnh đạo để thảo luận về cách các nhà lãnh đạo có thể chia sẻ thông tin giữa các vai trò, thích nghi với sự phát triển công nghệ nhanh chóng và xây dựng các cộng đồng chuyên nghiệp mạnh mẽ hơn. Thông qua các cuộc thảo luận, hội thảo và phiên họp tương tác, người tham dự sẽ xem xét cách hợp tác giữa các cấp độ kinh nghiệm và lĩnh vực có thể tăng cường việc ra quyết định và thúc đẩy đổi mới. Sự chuyển dịch sang lãnh đạo hợp tác được thúc đẩy bởi các yếu tố như chu kỳ phát triển công nghệ tăng tốc, nhu cầu xây dựng niềm tin công chúng và tỷ lệ lớn lực lượng lao động STEM đang đến tuổi nghỉ hưu. Hội nghị sẽ nhấn mạnh tầm quan trọng của một hệ sinh thái chung được xây dựng trên tinh thần hướng dẫn, học tập liên tục và chuyển giao kiến thức có chủ đích, nơi phát triển chuyên môn là một trao đổi đa hướng giữa các chuyên gia mới nổi, quản lý giữa sự nghiệp và các chuyên gia giàu kinh nghiệm.

7

Lovable được cho là đang đàm phán để tăng gấp đôi định giá lên 13,2 tỷ USD

Lovable reportedly in talks to double its valuation to $13.2B

TechCrunch AI Đọc bài viết →

Lovable, một công ty khởi nghiệp Thụy Điển chuyên về vibe-coding, được cho là đang đàm phán để huy động 300 triệu đô la với mức định giá 13,2 tỷ đô la. Điều này sẽ làm tăng hơn gấp đôi mức định giá của công ty từ 6,6 tỷ đô la đạt được vào tháng 12 năm ngoái. Menlo Ventures dự kiến sẽ dẫn đầu vòng này, theo một báo cáo. Lovable đã chứng kiến sự tăng trưởng đáng kể, đạt tốc độ doanh thu hàng năm là 500 triệu đô la vào tháng 6. Người dùng của công ty bao gồm các nhà sáng lập, nhà thiết kế và nhân viên bán hàng, cũng như các doanh nghiệp lớn như Workday, Asana và Nvidia. Vibe coding, cho phép người dùng xây dựng phần mềm bằng cách mô tả nó, là một trường hợp sử dụng phổ biến và có lợi nhuận cao cho AI. Vòng đầu tư này sẽ tiếp theo các công ty khởi nghiệp vibe-coding đáng chú ý khác, bao gồm Replit và Factory, những công ty cũng đã nhận được nguồn vốn đáng kể.

8

Viết lại Bun bằng Rust

Rewriting Bun in Rust

Simon Willison Đọc bài viết →

Jarred Sumner đã hoàn thành một bài đăng trên blog chi tiết về việc ông đã thành công trong việc viết lại dự án Bun bằng ngôn ngữ Rust, một ngôn ngữ mà ông đã chọn vì khả năng xử lý quản lý bộ nhớ hiệu quả hơn so với Zig, ngôn ngữ mà ông đã sử dụng trước đó. Quyết định viết lại Bun được thúc đẩy bởi số lượng lỗi cao của dự án, đặc biệt là lỗi use-after-free và double-free, những lỗi này là lỗi biên dịch trong Rust an toàn. Sumner đã sử dụng một bộ thử nghiệm được viết bằng TypeScript, phục vụ như một bộ thử nghiệm tuân thủ, cho phép một khung thử nghiệm tự động hóa nhiều phần của quá trình chuyển đổi ban đầu từ Bun sang Rust. Ông đã làm việc với một tác nhân mã hóa được hỗ trợ bởi một model LLM tiên phong, được gọi là Mythos/Fable, để hỗ trợ trong việc viết lại. Việc triển khai mới của Bun đã được triển khai trên Claude Code trong gần một tháng, với thời gian khởi động nhanh hơn 10% trên Linux. Dự án này chứng tỏ tiềm năng của việc sử dụng các tác nhân song song phối hợp để giải quyết các dự án đầy

> tham vọng và nhấn mạnh lợi ích của việc sử dụng Rust vì khả năng quản lý bộ nhớ của nó.

TIPS & TRICKS CHO DEV

Tối ưu hóa mã với GitHub Copilot

Vấn đề: Mã nguồn chưa tối ưu, cần cải thiện hiệu suất.
Cách làm: Sử dụng GitHub Copilot để đề xuất cải tiến mã, ví dụ "Optimize this function for performance".
Đánh giá: Hiệu quả trong việc cải thiện hiệu suất mã, nên dùng khi cần tối ưu hóa mã nguồn.

Tự động hoàn thành mã với Cursor

Vấn đề: Gãy dòng mã, cần tự động hoàn thành.
Cách làm: Sử dụng Cursor để hoàn thành mã tự động, ví dụ "Complete this line of code".
Đánh giá: Tiết kiệm thời gian, nên dùng khi cần hoàn thành mã nhanh chóng.

Hỗ trợ lập trình với Aider

Vấn đề: Cần hỗ trợ lập trình, tìm kiếm giải pháp.
Cách làm: Sử dụng Aider để tìm kiếm giải pháp, ví dụ "How to implement a sorting algorithm".
Đánh giá: Hiệu quả trong việc hỗ trợ lập trình, nên dùng khi cần tìm kiếm giải pháp lập trình.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
Dev cần biết cách tối ưu chi phí và hiệu năng của LLM (Large Language Model) để đảm bảo ứng dụng của mình hoạt động hiệu quả và tiết kiệm chi phí. Điều này đặc biệt quan trọng khi triển khai LLM trong các dự án thực tế.

2. Việc tối ưu hóa giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý của LLM, từ đó cải thiện trải nghiệm người dùng.

3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) để điều chỉnh LLM cho các use case cụ thể, giảm nhu cầu về tài nguyên tính toán.

4. Tip hoặc bước tiếp theo: Sử dụng các công cụ như Hugging Face Transformers để triển khai và tối ưu hóa LLM trong dự án của bạn.

Nguồn: Google News · Groq AI