BIS - MT · 30/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

"Trăm hay không bằng tay quen."

— Tục ngữ Việt Nam

Lý thuyết dù phong phú cũng không bằng thực hành thuần thục — kỹ năng chỉ đến từ luyện tập liên tục.

TIN TỨC NỔI BẬT

1

Giới thiệu Open Agent Specification (Agent Spec): Một cách biểu diễn thống nhất cho các AI Agent

Introducing the Open Agent Specification (Agent Spec): A Unified Representation for AI Agents

Oracle Blogs Đọc bài viết →

> Oracle đã giới thiệu Open Agent Specification (Agent Spec), một biểu diễn thống nhất cho các tác nhân AI. Agent Spec được thiết kế để cung cấp một khuôn khổ chung cho các nhà phát triển tạo và tương tác với các tác nhân AI trên nhiều nền tảng và ứng dụng khác nhau. Tiêu chuẩn này nhằm mục đích tiêu chuẩn hóa cách các tác nhân AI được biểu diễn, cho phép tích hợp và giao tiếp liền mạch giữa các hệ thống AI khác nhau. Agent Spec được xây dựng dựa trên các công nghệ hiện có như JSON-LD và JSON Schema, giúp nó dễ dàng tiếp cận và triển khai. Nó cung cấp một định dạng cấu trúc để biểu diễn các tác nhân AI, bao gồm khả năng, mục tiêu và hành vi của chúng. Biểu diễn tiêu chuẩn hóa này cho phép các nhà phát triển tạo ra các tác nhân AI có thể được hiểu và tương tác dễ dàng bởi các hệ thống khác, thúc đẩy khả năng tương tác và hợp tác. Agent Spec là một tiêu chuẩn mở, cho phép các nhà phát triển đóng góp và định hình sự phát triển của nó. Bằng cách cung cấp một biểu diễn thống nhất cho các tác nhân AI, Agent Spec có tiềm năng đẩy nhanh sự phát triển và áp dụng các công nghệ AI trên nhiều ngành công nghiệp và ứng dụng khác nhau.

2

Tăng tốc phát triển với Amazon Bedrock AgentCore MCP server | Trí tuệ nhân tạo

Accelerate development with the Amazon Bedrock AgentCore MCP server | Artificial Intelligence

Amazon Web Services (AWS) Đọc bài viết →

Amazon Web Services (AWS) đã giới thiệu máy chủ Amazon Bedrock AgentCore MCP, được thiết kế để tăng tốc phát triển trong lĩnh vực Trí tuệ Nhân tạo (AI). Máy chủ AgentCore MCP là một thành phần chính của nền tảng Amazon Bedrock, cung cấp một môi trường có thể mở rộng và bảo mật để xây dựng, đào tạo và triển khai các model AI. Máy chủ AgentCore MCP được tối ưu hóa cho tính toán hiệu suất cao, cho phép các nhà phát triển đào tạo nhanh chóng và hiệu quả các model AI phức tạp. Máy chủ này được xây dựng trên kiến trúc microservices, cho phép tích hợp liền mạch với các dịch vụ và công cụ AWS khác. Bằng cách tận dụng máy chủ AgentCore MCP, các nhà phát triển có thể tăng tốc quá trình phát triển, giảm chi phí và cải thiện chất lượng tổng thể của các model AI của họ. Nền tảng Amazon Bedrock được thiết kế để hỗ trợ một loạt các ứng dụng AI, từ xử lý ngôn ngữ tự nhiên đến tầm nhìn máy tính. Với máy chủ AgentCore MCP, các nhà phát triển có thể tận dụng toàn bộ tiềm năng của nền tảng, mở ra những khả năng mới cho đổi mới và tăng trưởng trong lĩnh vực AI.

3

Vai trò thầm lặng của Toán học và các thuật toán trong các hệ thống MCP & Multi-Agent

The Silent Role of Mathematics and Algorithms in MCP & Multi-Agent Systems

Cisco Blogs Đọc bài viết →

Trong một bài đăng gần đây trên Cisco Blogs, vai trò im lặng của toán học và thuật toán trong Nền tảng Đa đám mây (MCP) và Hệ thống Đa tác nhân được nhấn mạnh. Những hệ thống này rất quan trọng trong công nghệ hiện đại, cho phép giao tiếp và hợp tác liền mạch trên nhiều nền tảng khác nhau. Tuy nhiên, các khái niệm toán học và thuật toán cơ bản giúp chúng hoạt động thường không được chú ý. Toán học đóng vai trò quan trọng trong MCP, đảm bảo xử lý, lưu trữ và truy xuất dữ liệu hiệu quả. Nó cho phép phát triển các thuật toán phức tạp giúp phân tích dữ liệu, học máy và trí tuệ nhân tạo (AI). Những thuật toán này là thiết yếu trong Hệ thống Đa tác nhân, nơi nhiều tác nhân tự động tương tác và đưa ra quyết định dựa trên môi trường và mục tiêu của chúng. Bài đăng nhấn mạnh tầm quan trọng của nền tảng toán học trong MCP và Hệ thống Đa tác nhân, nhấn mạnh nhu cầu hiểu sâu hơn về những khái niệm này. Bằng cách nhận ra vai trò im lặng của

toán học và thuật toán, các nhà phát triển (developer) và nhà nghiên cứu có thể tạo ra các hệ thống hiệu quả, có thể mở rộng và bảo mật hơn, thúc đẩy đổi mới trong nhiều ngành công nghiệp, đặc biệt là trong lĩnh vực như API, LLM, và framework.

4

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 ít được biết đến đã có những đóng góp đáng kể, chẳng hạn như react-boilerplate và styled-components. Stoiber cũng đề cập đến sự phát triển của GitHub,□ đến cách Spectrum, một nền tảng ông đồng sáng lập, đã trở thành một phần của GitHub và giúp định hình GitHub Discussions. Ngoài ra, ông cũng nói về việc mua lại Stellate, một bộ nhớ đệm GraphQL, bởi Shopify và The Guild. Stoiber tin rằng các ứng dụng ChatGPT đại diện cho một bề mặt mới cho phát triển phần mềm, mang lại cơ hội đổi mới mới. Tập này được tài trợ bởi Coder.com, WorkOS, Notion, Fly.io và các đơn vị khác, cung cấp môi trường bảo mật, giải pháp xác thực và dịch vụ triển khai cho các developer.

5

Về việc sở hữu một codebase, và tại sao đó có thể là công việc khó khăn nhất trong ngành phần mềm

On owning a codebase, and why it may be the hardest job in software

Sourcegraph Blog Đọc bài viết →

Ngành công nghiệp phần mềm đang chứng kiến sự bùng nổ của các tác nhân mã hóa được hỗ trợ bởi AI, tạo ra lượng mã khổng lồ. Tuy nhiên, mặc dù có bước tiến công nghệ này, thế giới vẫn phụ thuộc nặng nề vào các cơ sở mã cũ đã có từ hàng thập kỷ. Những cơ sở mã này phức tạp và tinh vi, khiến cho việc sở hữu và hiểu chúng trở thành một nhiệm vụ đầy thách thức đối với các developer. Quy mô và tuổi thọ của các cơ sở mã này đặt ra những thách thức đáng kể, bao gồm việc giải mã các ngôn ngữ lập trình lỗi thời, tích hợp công nghệ mới và đảm bảo tính tương thích với các hệ thống hiện có. Kết quả là, việc sở hữu và duy trì các cơ sở mã này được coi là một trong những công việc khó

khăn nhất trong ngành công nghiệp phần mềm. Sự phức tạp và mong manh của các hệ thống cũ này đòi hỏi sự hiểu biết sâu sắc về hoạt động bên trong của chúng, khiến nó trở thành một nhiệm vụ đòi hỏi chuyên môn cao và đầy thách thức. Sự cần thiết của các chuyên gia có kỹ năng để điều hướng và duy trì các cơ sở mã này đang trở nên quan trọng hơn, nhấn mạnh tầm quan trọng của khía cạnh thường bị bỏ qua này trong phát triển phần mềm.

6

Các AI agent không phải là "đồng nghiệp" của bạn

AI agents are not your “coworkers”

MIT Tech Review Đọc bài viết →

Các nhà nghiên cứu đã phát hiện rằng việc tiếp thị các tác nhân AI như nhân viên kỹ thuật số có thể có những hậu quả không lường trước, bao gồm độ chính xác giảm và trách nhiệm giải trình giảm trong số công nhân. Một nghiên cứu của giáo sư kinh doanh Emma Wiles tại Đại học Boston đã phát hiện ra rằng khi các công cụ AI được gọi là "nhân viên", các nhà quản lý ít có khả năng phát hiện lỗi 18% và có nhiều khả năng chuyển trách nhiệm. Điều này là do việc định khung các tác nhân AI như đồng nghiệp có thể đảo ngược cảm giác trách nhiệm, khiến các nhà quản lý xem mình ít có trách nhiệm hơn đối với đầu ra của AI. Trên thực tế, gần một phần ba nhà quản lý trong nghiên cứu cho biết các công ty của họ đã định khung các tác nhân AI như nhân viên, với một số thậm chí còn liệt kê chúng trên biểu đồ tổ chức. Các chuyên gia cảnh báo rằng cách tiếp cận này có thể tạo ra những kỳ vọng không thực tế về khả năng của AI và tạo ra một văn hóa chuyển trách nhiệm, nơi các lỗi của con người được quy cho các tác nhân AI. Thay vào đó, các nhà nghiên cứu đề xuất rằng các công cụ AI nên được tối ưu hóa để cải thiện khả năng của con người, chứ không phải thay thế chúng. Một nghiên cứu gần đây tại Stanford đã phát hiện ra rằng công nhân thường có ý kiến khác nhau về những nhiệm vụ mà AI có thể giúp đỡ, và rằng tự động hóa có thể có lợi trong một số lĩnh vực, nhưng không phải trong các lĩnh vực khác.

7

Sau Orthogonality: Agency đạo đức đức hạnh và AI Alignment

After Orthogonality: Virtue-Ethical Agency and AI Alignment

The Gradient Đọc bài viết →

Bài viết này khám phá khái niệm về việc căn chỉnh trí tuệ nhân tạo (AI) với quyền lực và giá trị của con người. Tác giả cho rằng những người hợp lý không có mục tiêu, mà thay vào đó căn chỉnh hành động của họ với các thực hành, vốn là mạng lưới các hành động, khuynh hướng, tiêu chí đánh giá và tài nguyên. Để tạo ra các AI thực sự hỗ trợ quyền lực của con người, quá trình suy xét của các tác nhân AI phải chia sẻ một logic tương tự. Tác giả đề xuất rằng các khái niệm như minh bạch, hữu ích và vô hại là tự nhiên hơn cho các tác nhân giải thích chúng như các động lực trong mạng lưới này, thay vì như mục tiêu hoặc quy tắc. Bài viết cũng giới thiệu khái niệm về eudaimonia, hay sự thịnh vượng hợp lý và chủ động của con người, và cho rằng nó chỉ ra một cấu trúc suy xét khác với tính hợp lý kết quả tiêu chuẩn. Tác giả đề xuất tính hợp lý eudaimonic như một khuôn khổ hữu ích cho quyền lực và giá trị của AI được căn chỉnh với con người, với tiềm năng về sự ổn định và an toàn. Bài viết gợi ý rằng trực giác của chúng ta về sự thịnh vượng của con người ngụ ý rằng tính hợp lý eudaimonic là một hình thức tự nhiên và hiệu quả của quyền lực, và việc căn chỉnh AI với hình thức tính hợp lý này có thể giải quyết nhiều vấn đề an toàn AI cổ điển và nghịch lý.

8 Tính năng tạo ảnh AI cá nhân hóa của Gemini hiện đã miễn phí cho người dùng tại Mỹ

Gemini's personalized AI image generation is now free for US users

TechCrunch AI Đọc bài viết →

Google đã cung cấp tính năng tạo hình ảnh cá nhân hóa bằng AI, được hỗ trợ bởi Nano Banana, miễn phí cho tất cả người dùng đủ điều kiện tại Mỹ. Tính năng này, là một phần của ứng dụng Gemini, trước đây chỉ dành riêng cho người dùng Plus, Pro và Ultra. Nó sử dụng dữ liệu từ các kết nối tài khoản Google của người dùng, chẳng hạn như Gmail, Google Photos, YouTube và Tìm kiếm, để tạo ra hình ảnh dựa trên sở thích và ưu tiên duy nhất của họ. Người dùng có thể yêu cầu hình ảnh mà không cần chỉ định sở thích của họ, và Gemini cũng có thể kéo hình ảnh thực tế từ Google Photos. Tính năng này là tùy chọn, cho phép người dùng quyết định ứng dụng Gemini có thể truy cập vào ứng dụng nào. Sự mở rộng này diễn ra sau khi Google gần đây triển khai tính năng Trí tuệ Cá nhân hóa cho người dùng tại Ấn Độ và Nhật Bản. Ứng dụng Gemini đã vượt qua 750 triệu người dùng hoạt động hàng tháng, củng cố vị trí của nó trong không gian AI.

TIPS & TRICKS CHO DEV

Tạo Code Cơ Bản

Vấn đề: Tạo code cơ bản cho dự án mới tốn nhiều thời gian.
Cách làm: Sử dụng Claude Code với lệnh `claude init` để tạo code cơ bản. Ví dụ, với prompt "Tạo dự án Python mới", Claude Code sẽ tạo ra một dự án cơ bản.
Đánh giá: Hiệu quả cao, tiết kiệm thời gian, nhưng cần chỉnh sửa sau khi tạo.

Debug Code

Vấn đề: Debug code tốn nhiều thời gian và công sức.
Cách làm: Sử dụng Claude Code với lệnh `claude debug` để tìm và sửa lỗi. Ví dụ, với prompt "Sửa lỗi syntax trong file main.py", Claude Code sẽ giúp tìm và sửa lỗi.
Đánh giá: Hiệu quả cao, tiết kiệm thời gian, nhưng cần kiểm tra lại sau khi sửa.

Refactor Code

Vấn đề: Refactor code để cải thiện hiệu suất và dễ bảo trì.
Cách làm: Sử dụng Claude Code với lệnh `claude refactor` để cải thiện code. Ví dụ, với prompt "Cải thiện hiệu suất của hàm tính toán", Claude Code sẽ giúp refactor code.
Đánh giá: Hiệu quả cao, cải thiện hiệu suất, nhưng cần kiểm tra lại sau khi refactor.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Các nhà phát triển cần biết cách tối ưu chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM) để đảm bảo ứng dụng AI của họ hoạt động hiệu quả và tiết kiệm. Điều này đặc biệt quan trọng khi triển khai LLM trong các ứng dụng thực tế. Việc tối ưu hóa giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý.
3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) cho phép điều chỉnh LLM cho các trường hợp sử dụng cụ thể mà không cần phải đào tạo lại toàn bộ mô hình, giúp tiết kiệm tài nguyên và thời gian.
4. Tip hoặc bước tiếp theo: Hãy xem xét việc áp dụng các kỹ thuật như quantization, pruning, và knowledge distillation để tối ưu hóa thêm LLM và đạt được hiệu suất tốt hơn trong ứng dụng của bạn.