BIS - MT · 03/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“We must accept finite disappointment, but never lose infinite hope.”

↪ Chúng ta phải chấp nhận thất vọng hữu hạn, nhưng không bao giờ mất đi hy vọng vô hạn.

— Martin Luther King Jr.

Giữ hy vọng không có nghĩa là ngây thơ — đó là sự lựa chọn tích cực để tiếp tục tiến về phía trước dù có gặp thất bại tạm thời.

TIN TỨC NỔI BẬT

1

Hướng dẫn hoàn chỉnh về tệp 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 này cung cấp một hướng dẫn chi tiết về các tệp tin bộ nhớ của tác nhân AI, tập trung cụ thể vào CLAUDE.md và AGENTS.md. Tác giả đi sâu vào thế giới của các tác nhân AI, giải thích tầm quan trọng của các tệp tin bộ nhớ trong hoạt động của chúng. CLAUDE.md được mô tả là một tệp tin quan trọng chứa siêu dữ liệu về tác nhân AI, bao gồm tên, mô tả và các chi tiết liên quan khác. AGENTS.md, mặt khác, là một tệp tin lưu trữ thông tin về các tác nhân AI, chẳng hạn như khả năng và giới hạn của chúng. Bài viết cũng đề cập đến các tệp tin bộ nhớ của tác nhân AI khác, mặc dù các chi tiết cụ thể không được cung cấp. Nó nhấn mạnh tầm quan trọng của các tệp tin này trong việc hiểu và quản lý các tác nhân AI, đang ngày càng được sử dụng trong nhiều ứng dụng, bao gồm cả chatbot và trợ lý ảo. Bằng cách khám phá các tệp tin bộ nhớ này, các nhà phát triển có thể thu được những hiểu biết quý giá về hoạt động nội bộ của các tác nhân AI và cải thiện hiệu suất và chức năng của chúng. Bài viết này phục vụ như một tài nguyên toàn diện cho những người quan tâm đến việc phát triển và quản lý tác nhân AI, đặc biệt là trong việc xây dựng và tích hợp các mô hình LLM, API và framework khác nhau để tạo ra các giải pháp AI hiệu quả hơn.

2 Lớp lệnh AI nhà kho đa agent cho phép xuất sắc hoạt động 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 giới thiệu Lớp lệnh AI Nhà kho Đa tác nhân, một công nghệ tiên tiến được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng. Giải pháp đổi mới này tận dụng AI và học máy để tối ưu hóa hoạt động nhà kho, cho phép các doanh nghiệp đưa ra quyết định dựa trên dữ liệu. Lớp lệnh AI Nhà kho Đa tác nhân là một trung tâm điều hành tích hợp các tác nhân AI khác nhau để quản lý và điều phối các hoạt động nhà kho. Các tác nhân này làm việc cùng nhau để phân tích dữ liệu từ các nguồn khác nhau, bao gồm mức tồn kho, lịch trình giao hàng và hiệu suất thiết bị. Bằng cách làm như vậy, hệ thống có thể xác định các điểm nghẽn, dự đoán các vấn đề tiềm ẩn và đề xuất cải tiến để tối ưu hóa hoạt động nhà kho. Công nghệ này cũng cung cấp khả năng hiển thị thời gian thực vào hoạt động chuỗi cung ứng, cho phép các doanh nghiệp theo dõi chuyển động hàng tồn kho, trạng thái giao hàng và các chỉ số chính khác. Mức độ minh bạch này cho phép các công ty phản ứng nhanh chóng với những thay đổi về nhu cầu, giảm chi phí và cải thiện sự hài lòng của khách hàng. Bằng cách tận dụng sức mạnh của AI và học máy, Lớp lệnh AI Nhà kho Đa tác nhân có tiềm năng cách mạng hóa cách các doanh nghiệp quản lý nhà kho và chuỗi cung ứng của họ.

3 Nhà máy Agent: Kết nối các agent, ứng dụng và dữ liệu với các tiêu chuẩn mở mới như MCP và A2A

Agent Factory: Connecting agents, apps, and data with new open standards like MCP and A2A

Microsoft Azure Đọc bài viết →

Microsoft đã giới thiệu Agent Factory, một sáng kiến mới nhằm kết nối các tác nhân, ứng dụng và dữ liệu thông qua các tiêu chuẩn mở. Dự án này sử dụng Cloud PC (MCP) và giao thức Ứng dụng đến Ứng dụng (A2A) của Microsoft để tạo điều kiện cho sự tương tác liền mạch giữa các thành phần khác nhau. Agent Factory nhằm mục đích đơn giản hóa quá trình tích hợp bằng cách cung cấp một khuôn khổ tiêu chuẩn hóa cho giao tiếp và trao đổi dữ liệu. Điều này cho phép các nhà phát triển xây dựng các ứng dụng hiệu quả và có khả năng mở rộng hơn, có thể dễ dàng tương tác với các hệ thống và dịch vụ khác. Việc sử dụng

các tiêu chuẩn mở như MCP và A2A cho phép có sự linh hoạt và khả năng tương tác cao hơn, làm cho việc các hệ thống khác nhau làm việc cùng nhau trở nên dễ dàng hơn. Điều này có thể dẫn đến năng suất lao động được cải thiện, chi phí giảm và trải nghiệm người dùng tổng thể được nâng cao. Bằng cách tận dụng các tiêu chuẩn mở này, Agent Factory có tiềm năng cách mạng hóa cách các ứng dụng và dữ liệu tương tác, mở đường cho các giải pháp sáng tạo và kết nối hơn.

4

Tại sao tương lai của AI agent lại xoay quanh việc điều khiển

Why the future of agentic AI is all about the harness

BD Tech Talks Đọc bài viết →

Tương lai của AI agentic, một loại trí tuệ nhân tạo có thể thực hiện hành động và đưa ra quyết định, đang thay đổi từ việc chỉ tập trung vào xây dựng các model lớn hơn và cung cấp cho chúng nhiều dữ liệu hơn. Theo một bài báo mới từ UC Berkeley, điểm nghẽn chính tiếp theo trong AI agentic là "system scaling", hoặc việc mở rộng "harness" giúp dịch các câu trả lời của model thành hành vi trong thế giới thực. Harness này là một hệ thống phức tạp bao gồm nhiều thành phần, chẳng hạn như lưu trữ bộ nhớ, xây dựng ngữ cảnh, định tuyến kỹ năng và xác minh và quản trị. Các thành phần này hoạt động cùng nhau để quản lý hành vi của agent và đảm bảo hiệu suất tối ưu. Các nhà nghiên cứu hiện đang nhận ra rằng toàn bộ hệ thống, bao gồm model và cơ sở hạ tầng xung quanh, phải được tối ưu hóa cho hiệu suất. Các framework agent hiện đại hoạt động như cơ sở hạ tầng hệ thống mạnh mẽ, chứ không chỉ là các wrapper prompt đơn giản. Các tác giả đề xuất chia hệ thống AI thành ba lớp chức năng: prompt, kỹ năng và bộ nhớ. Prompt định nghĩa các mục tiêu ngay lập tức, kỹ năng là các mẫu thực hiện có thể tái sử dụng và bộ nhớ lưu giữ các thực qua các phiên. Tuy nhiên, mỗi lớp này lại đưa ra những thách thức mới, chẳng hạn như định tuyến, thành phần và ủy quyền, và suy giảm bộ nhớ, ô nhiễm và tổng quát hóa quá mức.

5

Cyera hướng tới định giá 12 tỷ USD với hệ số ARR 80 lần bất chấp lỗ hoạt động

Cyera eyes $12B valuation at 80x ARR multiple despite operating losses

TechCrunch AI Đọc bài viết →

Cyera, một công ty bảo mật lưu trữ dữ liệu, được báo cáo đang hoàn thiện vòng tài trợ do Evolution Equity Partners dẫn đầu, định giá công ty ở mức 12 tỷ đô la. Định giá này dựa trên doanh thu định kỳ hàng năm (ARR) là 150 triệu đô la, đưa công ty đến hệ số ARR là 80, cao hơn nhiều công ty khởi nghiệp AI tăng trưởng nhanh. Mặc dù có sự tăng trưởng doanh thu đáng kể, Cyera không có lợi nhuận và đã hoạt động trong tình trạng thua lỗ, với một số chi phí được chỉ đạo để thuê nhân viên bán hàng. Công ty đã thêm 500 việc làm từ đầu năm đến nay. Vòng tài trợ mới này dự kiến sẽ đưa tổng số vốn của Cyera lên ít nhất 2 tỷ đô la, sau vòng Series F trước đó là 400 triệu đô la vào năm 2025.

6 Qwen3.7-Plus của Alibaba hỗ trợ nhập văn bản, video và hình ảnh với chi phí thấp 0,4/1,6 USD cho mỗi 1 triệu token — nhưng nó là độc quyền

Alibaba's Qwen3.7-Plus supports text, video and imagery inputs at low cost of $0.4/$1.6 per 1M token — but it's proprietary

VentureBeat Đọc bài viết →

Alibaba đã phát hành Qwen3.7-Plus, một mô hình AI ngôn ngữ lớn mới hỗ trợ đầu vào văn bản, video và hình ảnh với chi phí thấp hơn so với người tiền nhiệm của nó, Qwen3.7-Max. Mô hình này có sẵn dưới giấy phép độc quyền và có thể được truy cập thông qua Cloud Model Studio của Alibaba. Qwen3.7-Plus tự hào có chi phí thấp hơn 60% so với Qwen3.7-Max, với giá 0,4 đô la cho 1 triệu token cho đầu vào và 1,6 đô la cho 1 triệu token cho đầu ra. Mô hình này có khả năng đa phương thức, cho phép nó tạo ra hình ảnh cấp doanh nghiệp và phân tích video, hình ảnh và ảnh chụp màn hình. Nó cũng có một cửa sổ ngữ cảnh 1 triệu token và phân bổ lên đến 256K token cho quá trình xử lý chuỗi suy nghĩ nội bộ. Tính năng này, được gọi là "preserve_thinking", cho phép mô hình giữ lại các vòng lặp logic nội bộ và ngăn chặn mất ngữ cảnh trong các nhiệm vụ phức tạp. Qwen3.7-Plus đã được thử nghiệm so với các mô hình khác, thể hiện hiệu suất cạnh tranh trên các nhiệm vụ đa phương thức và tác nhân. Tuy nhiên, nó vẫn thấp hơn một số mô hình hàng đầu về các chỉ số khả năng thô. Mô hình này được thiết kế để thay thế các mô hình tiền phong trong các công việc của nhà phát triển tần suất cao, tự động hóa quy trình robot và đường ống kỹ thuật dữ liệu. Alibaba đã cấu trúc việc phân phối API của mình để phù hợp với các khuôn khổ doanh nghiệp mã nguồn mở và độc quyền hiện có, khiến nó dễ dàng tích hợp vào các ngăn xếp công nghệ hiện có. Mô hình này cũng cung cấp các điểm giá

bộ nhớ đệm chi tiết, khiến các lần lặp lại của tác nhân đa lượt tần suất cao trở nên thực tế về mặt kinh tế ở quy mô doanh nghiệp. Tuy nhiên, mô hình này chỉ có sẵn dưới giấy phép độc quyền, điều này gây ra lo ngại về tuân thủ và chủ quyền cho các doanh nghiệp hoạt động dưới các nghĩa vụ cư trú dữ liệu nghiêm ngặt. Mặc dù vậy, Qwen3.7-Plus trình bày một giải pháp thay thế thực tế cho người mua doanh nghiệp đang tìm cách xây dựng các vòng lặp phần mềm tự động, có khả năng hình ảnh mạnh mẽ mà không vượt quá ngân sách suy luận của họ.

7 Travelers triển khai yêu cầu bồi thường bằng AI trên toàn quốc với OpenAI

Travelers deploys AI-powered claims countrywide with OpenAI

OpenAI Blog Đọc bài viết →

Travelers, nhà cung cấp bảo hiểm hàng đầu, đã giới thiệu Trợ lý Yêu cầu Bảo hiểm được hỗ trợ bởi AI trên toàn quốc. Công cụ đổi mới này, được phát triển hợp tác với OpenAI, nhằm mục đích đơn giản hóa quá trình nộp yêu cầu bồi thường cho khách hàng. Trợ lý hỗ trợ bởi AI cung cấp hỗ trợ 24/7, cho phép khách hàng tìm kiếm sự hỗ trợ tại thời điểm thuận tiện cho họ. Một trong những lợi ích chính của hệ thống mới này là khả năng mở rộng hoạt động trong thời kỳ nhu cầu cao điểm. Điều này có nghĩa là Travelers có thể quản lý hiệu quả một lượng lớn yêu cầu bồi thường mà không ảnh hưởng đến chất lượng dịch vụ hoặc thời gian phản hồi. Bằng cách tận dụng công nghệ AI, công ty có thể cung cấp trải nghiệm được sắp xếp hợp lý và phản hồi hơn cho khách hàng của mình. Việc triển khai Trợ lý Yêu cầu Bảo hiểm được hỗ trợ bởi AI đánh dấu một bước tiến quan trọng trong nỗ lực của Travelers nhằm nâng cao sự hài lòng của khách hàng và hiệu quả hoạt động. Với công cụ mới này, khách hàng có thể mong đợi một trải nghiệm yêu cầu bồi thường liền mạch và không phức tạp hơn.

8 Tải xuống: AI có thể chạy bộ phận hành chính của bạn bây giờ

The Download: AI can run your admin department now

MIT Tech Review Đọc bài viết →

Phiên bản hôm nay của The Download nhấn mạnh khả năng ngày càng tăng của trí tuệ nhân tạo (AI) trong các nhiệm vụ hành chính. Các công ty lớn thường có nguồn lực để thuê chuyên gia cho các chức

> năng kinh doanh khác nhau, nhưng các doanh nghiệp nhỏ có thể không có cùng sự sang trọng. Các mô hình AI hiện có thể thực hiện công việc hành chính cơ bản, bao gồm tổ chức ghi chú, tóm tắt cuộc họp, lập hóa đơn, đặt mục tiêu và lập kế hoạch truyền thông xã hội. Sự phát triển này có thể là một yếu tố thay đổi cuộc chơi cho các doanh nghiệp nhỏ, cho phép họ đưa AI vào công việc và tối ưu hóa hoạt động của mình. Trong các tin tức công nghệ khác, Anthropic đã nộp đơn xin chào bán công khai ban đầu (IPO) một cách bí mật trước OpenAI, với công ty nhằm mục đích niêm yết công khai sớm nhất là vào mùa thu này. Liên minh Châu Âu (EU) cũng có thể loại trừ các gã khổng lồ đám mây của Mỹ khỏi các hợp đồng quan trọng, giảm sự phụ thuộc vào công nghệ Mỹ. Ngoài ra, Florida đã trở thành tiểu bang đầu tiên kiện OpenAI vì những rủi ro an toàn cho trẻ em được cho là trong ứng dụng trò chuyện ChatGPT của họ.

TIPS & TRICKS CHO DEV

Chain-of-Thought Prompting

Vấn đề: Dev gặp khó khăn khi tạo prompt cho các task phức tạp.
Cách làm: Sử dụng chain-of-thought prompting, ví dụ "Break down the solution into smaller steps" để tạo ra chuỗi suy nghĩ logic.
Đánh giá: Hiệu quả cao cho các task phức tạp, nhưng có thể không phù hợp cho task đơn giản.

Few-Shot Learning

Vấn đề: Dev cần train mô hình với dữ liệu hạn chế.
Cách làm: Sử dụng few-shot learning, ví dụ "Train a model to classify images with only 5 examples per class" để tận dụng dữ liệu ít.
Đánh giá: Hiệu quả cao cho các mô hình nhỏ, nhưng có thể không phù hợp cho mô hình lớn.

System Prompt Design

Vấn đề: Dev cần thiết kế prompt cho hệ thống phức tạp.
Cách làm: Sử dụng system prompt design, ví dụ "Design a prompt to extract information from a database" để tạo ra prompt hiệu quả.
Đánh giá: Hiệu quả cao cho các hệ thống phức tạp, nhưng yêu cầu dev có kinh nghiệm về thiết kế prompt.

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 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 đặc biệt quan trọng khi tích hợp AI vào ứng dụng, vì LLM có thể tiêu thụ nhiều tài nguyên và ảnh hưởng đến hiệu suất của hệ thống. Việc tối ưu hóa LLM giúp giảm thiểu chi phí và 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 mô hình cho phù hợp với từng trường hợp cụ thể, giảm kích thước mô hình và tăng tốc độ xử lý.

4. Tip hoặc bước tiếp theo: Nên xem xét sử dụng các công cụ như Hugging Face Transformers hoặc TensorFlow để tối ưu hóa LLM và cải thiện hiệu suất của ứng dụng.