BIS - MT · 15/08/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“The future belongs to those who believe in the beauty of their dreams.”

↪ Tương lai thuộc về những người tin vào vẻ đẹp của ước mơ mình.

— Eleanor Roosevelt

Niềm tin vào ước mơ là bước đầu tiên để biến nó thành hiện thực — không có giấc mơ nào quá lớn nếu bạn thực sự tin.

TIN TỨC NỔI BẬT

1

Tình hình các Agent Framework: Lựa chọn Runtime phù hợp cho việc triển khai AI cấp doanh nghiệp

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

Medium Đọc bài viết →

Bài viết "State of Agent Frameworks: Choosing the Right Runtime for Enterprise AI Execution" cung cấp cái nhìn tổng quan về tình hình hiện tại của các agent framework trong việc triển khai AI cấp doanh nghiệp. Các agent framework là những nền tảng phần mềm cho phép phát triển và triển khai các agent thông minh, vốn là những chương trình phần mềm có khả năng nhận thức, suy luận và hành động trong các môi trường phức tạp. Bài viết nhấn mạnh tầm quan trọng của việc lựa chọn runtime phù hợp cho việc triển khai AI cấp doanh nghiệp, vì nó có thể tác động đáng kể đến hiệu năng, khả năng mở rộng và khả năng bảo trì của các hệ thống AI. Bài viết thảo luận về nhiều agent framework khác nhau, bao gồm OpenCog, PyTorch, TensorFlow và Rasa, mỗi framework đều có những điểm mạnh và điểm yếu riêng. Bài viết cũng đề cập đến những thách thức trong việc tích hợp các hệ thống AI với hạ tầng doanh nghiệp hiện có, như hệ thống lưu trữ dữ liệu và bảo mật. Nó nhấn mạnh sự cần thiết của một agent framework linh hoạt và dễ thích nghi, có thể tích hợp liền mạch với nhiều hệ thống doanh nghiệp khác nhau. Cuối cùng, bài viết kết luận rằng việc lựa chọn đúng agent framework là rất quan trọng để triển khai AI cấp doanh nghiệp thành công, và việc cân nhắc kỹ lưỡng các yếu tố như hiệu năng, khả năng mở rộng và tích hợp là điều cần thiết.

2 AWS mở mã nguồn một MCP Server cho Bedrock AgentCore để tinh gọn quá trình phát triển AI Agent

AWS Open-Sources an MCP Server for Bedrock AgentCore to Streamline AI Agent Development

MarkTechPost Đọc bài viết →

Amazon Web Services (AWS) đã mở mã nguồn một MCP server để hỗ trợ Bedrock AgentCore. Động thái này nhằm mục đích đơn giản hóa việc phát triển các AI agent. MCP server là một thành phần cốt lõi của Bedrock AgentCore, một framework để xây dựng và triển khai các AI agent. Bằng cách mở mã nguồn MCP server, AWS đang cung cấp cho cộng đồng developer một nền tảng chuẩn hóa và mở để phát triển AI agent. Điều này có thể dẫn đến tăng cường hợp tác, đổi mới và hiệu quả trong lĩnh vực nghiên cứu và phát triển AI. Bedrock AgentCore được thiết kế để cung cấp một interface thống nhất cho việc xây dựng và triển khai các AI agent trên nhiều nền tảng và môi trường khác nhau. Việc mở mã nguồn MCP server được kỳ vọng sẽ đẩy nhanh quá trình phát triển AI agent và cho phép các nhà nghiên cứu, developer tập trung vào việc tạo ra các AI model tinh vi và hiệu quả hơn.

3 Agent Factory: Kết nối các agent, ứng dụng và dữ liệu bằng các open standard 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 agent, ứng dụng và dữ liệu thông qua các open standard. Dự án này sử dụng các protocol Cloud Computing Platform (MCP) và Application-to-Application (A2A) của Microsoft để tạo điều kiện tích hợp và giao tiếp liền mạch giữa các hệ thống khác nhau. Agent Factory tìm cách thu hẹp khoảng cách giữa các ứng dụng khác nhau, cho phép chúng hoạt động cùng nhau hiệu quả và năng suất hơn. Bằng cách tận dụng MCP và A2A, các developer có thể tạo ra các giải pháp mạnh mẽ và có khả năng mở rộng hơn, có thể tương tác với nhiều nguồn dữ liệu và service. Việc giới thiệu Agent Factory được kỳ vọng sẽ nâng cao trải nghiệm người dùng tổng thể bằng cách cung cấp một môi trường gắn kết và tinh gọn hơn. Bằng cách giảm bớt sự phức tạp và rào cản liên quan đến việc tích hợp các hệ thống khác nhau, Agent Factory có tiềm năng thúc đẩy đổi mới và tăng trưởng trong nhiều ngành công nghiệp. Cam kết của Microsoft đối với các

open standard với Agent Factory có khả năng thúc đẩy sự hợp tác và khả năng tương tác lớn hơn giữa các developer và tổ chức.

4

AI ưu tiên tuân thủ: Chứng minh agent provenance cho các đội ngũ engineering chịu quy định

Compliance-first AI: proving agent provenance for regulated engineering teams

Sourcegraph Blog Đọc bài viết →

Các đội ngũ engineering chịu quy định đối mặt với thách thức đáng kể trong việc áp dụng AI agent do lo ngại về trách nhiệm giải trình và tuân thủ. Mặc dù AI agent có thể viết code, vấn đề nằm ở việc chứng minh rằng chúng đã truy cập đúng thông tin trước khi thực hiện thay đổi. Điều này được gọi là agent provenance, đề cập đến bằng chứng về những file mà một AI agent đã đọc và lý do tại sao. Một audit trail rõ ràng là chưa đủ, vì một agent có thể ghi lại lý do của nó mà không thực sự truy cập thông tin cần thiết. Để giải quyết vấn đề này, cần có một bản ghi minh bạch về các file mà agent đã đọc và lý do tại sao. Điều này cho phép các auditor xác minh tính chính xác và đầy đủ của logic của agent và xác nhận rằng nó đã hành động dựa trên bộ yêu cầu đầy đủ chỉ sử dụng context code liên quan và được ủy quyền. Ngược lại, các human engineer có thể giải thích hành động và quyết định của họ thông qua các bản ghi Pull Request và phê duyệt. Khoảng cách giữa phát triển của con người và AI agent là đáng kể, vì AI agent thiếu khả năng cung cấp context và giải thích cho hành động của

5

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 Tổng quát Nhân tạo (AGI) đang sắp xảy ra. Tuy nhiên, giả định này có thể là sai lầm. Sự thành công của những mô hình này có thể được quy cho khả năng của chúng trong việc mở rộng hiệu quả trên phần cứng hiện có, chứ không phải là một giải pháp sâu sắc cho vấn đề về trí tuệ. Phương pháp đa mô thức, bao gồm việc kết hợp nhiều mô thức khác nhau để đạt được trí tuệ tổng quát, không thể dẫn đến AGI ở mức độ con người trong thời gian gần. Phương pháp này tập trung vào việc vá các mô thức khác nhau lại với nhau, thay vì coi việc thể hiện và tương tác với môi trường là những khía cạnh chính

của trí tuệ. Một AGI thực sự phải có khả năng giải quyết các vấn đề phát sinh từ thực tế vật lý, chẳng hạn như sửa chữa một chiếc xe hơi hoặc tháo một nút thắt. Để đạt được điều này, một hình thức trí tuệ cơ bản nằm trong một mô hình thế giới vật lý là cần thiết. Các Mô hình Ngôn ngữ Lớn (LLM) hiện tại có thể không sở hữu mức độ hiểu biết này, vì chúng có nhiều khả năng học các quy tắc nông để dự đoán token thay vì một sự hiểu biết sâu sắc về thực tế. Điều này đã dẫn đến sự nhầm lẫn về ý nghĩa của việc hiểu ngôn ngữ và thế giới.

6

Đừng phân loại. Hãy "hallucinate"!

Don't classify. Hallucinate!

Simon Willison Đọc bài viết →

Một chuyên gia công nghệ đã đề xuất một giải pháp sáng tạo cho thách thức phân loại nội dung sử dụng Large Language Models (LLMs). Vấn đề này phát sinh khi nội dung hiện có có số lượng thẻ từ (tag) lớn, khiến cho model khó xác định được các thẻ từ phù hợp. Để vượt qua điều này, chuyên gia đề xuất cho phép model "hoang tưởng" hoặc tạo ra các thẻ từ mới mà không tham khảo từ vựng hiện có. Sau đó, model có thể sử dụng vector embeddings để khớp các thẻ từ được tạo ra với những thẻ từ tương tự nhất trong tập dữ liệu hiện có. Phương pháp này có thể được cải thiện bằng cách cung cấp cho model các ví dụ về cấu trúc thẻ từ mong muốn, chẳng hạn như danh mục và sub-danh mục. Phương pháp này đã được chứng minh là hiệu quả trong việc tạo ra các phân loại mới và chính xác.

7

Tình hình các Open Model: Những quan sát mùa hè 2026

State of Open Models: Summer 2026 Observations

Hugging Face Blog Đọc bài viết →

Cảnh quan mô hình AI đã trải qua những thay đổi đáng kể giữa tháng 1 và tháng 8 năm 2026. Theo một báo cáo gần đây, số lượng mô hình công khai trên trung tâm HF đã tăng, với 2,96 triệu mô hình và 1 triệu tập dữ liệu có sẵn. Tuy nhiên, phân phối vẫn bị lệch, với 85,6% mô hình có ít hơn 200 lần tải xuống trong suốt thời gian tồn tại. Các phòng thí nghiệm Trung Quốc đã tạo ra tác động đáng kể, phát hành các mô hình lớn và hiệu suất cao vượt qua các mô hình từ các phòng thí nghiệm Mỹ. Trên thực tế, các mô hình lớn nhất từ các phòng thí nghiệm Trung Quốc đã có từ 754B đến 2,78 nghìn tỷ tham số, trong

khi các phòng thí nghiệm Mỹ thường ở dưới 130B tham số. Sự thay đổi này đã dẫn đến sự thay đổi trong hồ sơ kích thước của mô hình, nơi xây dựng mô hình lớn không còn là yếu tố phân biệt, mà thay vào đó là một tuyên bố về ý định. Báo cáo cũng lưu ý rằng các nhà cung cấp phần cứng, chẳng hạn như AMD và NVIDIA, hiện là những người chơi chính trong không gian AI mã nguồn mở, phát hành hơn 200 kho mô hình mới mỗi. Sự thay đổi này đã dẫn đến sự giảm số lượng phát hành mô hình mới từ các phòng thí nghiệm AI truyền thống, chẳng hạn như Google và Meta. Báo cáo kết luận rằng trọng tâm của AI mã nguồn mở đã chuyển từ các phòng thí nghiệm mô hình sang các công ty phần cứng và cơ sở hạ tầng.

8

Chi phí ẩn của HDMI: Giải mã bí ẩn nhãn Ultra96

The Hidden Cost of HDMI: Unraveling the Mystery of the Ultra96 Label

Dev.to AI Đọc bài viết →

Thế giới công nghệ đầy rẫy những từ viết tắt và thuật ngữ chuyên môn, khiến nhiều người tiêu dùng cảm thấy bối rối. Một thuật ngữ như vậy là nhãn "Ultra96", một dấu chứng nhận được sử dụng bởi Cơ quan Cấp phép HDMI để chỉ rằng một sản phẩm đáp ứng các tiêu chuẩn nhất định cho việc truyền dữ liệu tốc độ cao. Nhãn này thường được tìm thấy trên cáp HDMI, bộ chuyển đổi và các thiết bị khác hỗ trợ truyền video và âm thanh độ nét cao. Nhãn Ultra96 là một tập con của chương trình chứng nhận HDMI, đảm bảo sản phẩm đáp ứng các tiêu chuẩn cụ thể về hiệu suất, chất lượng và độ tin cậy. Tiền tố "Ultra" chỉ rằng sản phẩm đã được thử nghiệm và chứng nhận để đáp ứng các tiêu chuẩn cao nhất cho việc truyền dữ liệu tốc độ cao, thường trên 10 Gbps. Nhãn này thường được sử dụng như một công cụ tiếp thị để chứng minh giá cao hơn, ngay cả khi sản phẩm không nhất thiết mang lại hiệu suất tốt hơn. Kết quả là, nhiều người tiêu dùng đang trả quá nhiều tiền cho cáp HDMI và bộ chuyển đổi không đáp ứng tiêu chuẩn Ultra96.

TIPS & TRICKS CHO DEV

Cài đặt Ollama

Vấn đề: Phải chạy LLM trên máy cá nhân mà không cần internet.

Cách làm: Sử dụng Ollama, chạy lệnh `ollama install` để cài đặt, sau đó `ollama run` để chạy mô hình.

Đánh giá: Hiệu quả cao, phù hợp khi làm việc offline.

Tối ưu LM Studio

Vấn đề: LM Studio chạy chậm trên máy yếu.

Cách làm: Sử dụng lệnh `lm-studio --optimize` để tối ưu hóa, giảm thiểu tài nguyên.

Đánh giá: Nhanh hơn, tiết kiệm RAM.

Chạy Prompt

Vấn đề: Cần chạy thử prompt trên LLM.

Cách làm: Sử dụng lệnh `ollama prompt "Đây là câu hỏi thử nghiệm"` để chạy thử.

Đánh giá: Hiệu quả, giúp kiểm tra mô hình nhanh.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Dev cần biết cách tối ưu chi phí và hiệu năng LLM để giảm thiểu chi phí vận hành và cải thiện hiệu suất của ứng dụng. Điều này giúp đảm bảo rằng mô hình AI hoạt động hiệu quả và không tiêu tốn quá nhiều tài nguyên. Việc tối ưu hóa cũng giúp cải thiện thời gian phản hồi và tăng trải nghiệm người dùng.
3. Ví dụ, có thể sử dụng kỹ thuật **fine-tuning** và **LoRA** để điều chỉnh mô hình LLM cho phù hợp với use case cụ thể, giảm thiểu số lượng tham số và cải thiện hiệu suất.
4. Tip hoặc bước tiếp theo: Nghiên cứu và áp dụng các kỹ thuật tối ưu hóa LLM như **quantization**, **pruning** và **knowledge distillation** để cải thiện hiệu suất và giảm chi phí vận hành.