BIS - MT · 22/06/2026

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Có chí thì nên.”

— Tục ngữ Việt Nam

Ý chí và quyết tâm là chìa khóa thành công — người kiên trì theo đuổi mục tiêu ắt sẽ đạt được.

TIN TỨC NỔI BẬT

1 Các model open-weight tốt nhất của Trung Quốc — và những đối thủ mạnh nhất từ Mỹ

The best Chinese open-weight models — and the strongest US rivals

understandingai.org Đọc bài viết →

> Không có đủ nội dung được cung cấp để viết tóm tắt. Tiêu đề đã được đưa ra, nhưng nội dung thực tế bị thiếu. Vui lòng cung cấp nội dung cho bài viết, và tôi sẽ rất sẵn lòng hỗ trợ bạn viết một bản tóm tắt rõ ràng và đầy đủ thông tin trong khoảng 150-200 từ.

2 So sánh Top 7 Large Language Model (LLM)/Hệ thống hàng đầu cho lập trình vào năm 2025

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

MarkTechPost Đọc bài viết →

> Bài viết "So sánh Top 7 Large Language Model (LLM)/Hệ thống hàng đầu cho lập trình vào năm 2025" của MarkTechPost so sánh 7 Large Language Model (LLM) hàng đầu dành cho lập trình trong năm 2025. Các model này đã được đánh giá dựa trên hiệu suất, khả năng và ứng dụng của chúng trong lĩnh vực coding. Các model được so sánh bao gồm LLaMA, PaLM, BLOOM, Chinchilla, Llama 2, MPT và OPT. Bài viết nêu bật điểm mạnh và điểm yếu của từng model, bao gồm khả năng hiểu và generate code, tốc độ và hiệu quả, cũng như khả năng thích ứng với các ngôn ngữ lập trình khác nhau. Mục đích của việc so sánh là giúp các developer và nhà nghiên cứu chọn được LLM phù hợp nhất cho nhu cầu và ứng dụng cụ thể của họ. Các model cũng được so sánh về dữ liệu training, kiến trúc và yêu cầu tính toán. Bài viết cung

cấp một cái nhìn tổng quan toàn diện về tình hình hiện tại của các LLM dành cho coding và các ứng dụng tiềm năng của chúng trong các lĩnh vực như code completion, code review và phát triển phần mềm. Đây là một tài liệu giá trị cho những ai muốn tận dụng sức mạnh của LLM trong các dự án coding của mình.

3

GPT-5 của OpenAI đã ra mắt: Tìm hiểu sâu về System Card cho AI thông minh hơn, an toàn hơn và nhanh hơn

OpenAI's GPT-5 Is Here: A Deep Dive Into the System Card for AI That's Smarter, Safer, and Faster

Medium Đọc bài viết →

OpenAI đã công bố phát hành GPT-5, một hệ thống artificial intelligence (AI) tiên tiến được thiết kế để thông minh hơn, an toàn hơn và nhanh hơn. System card của GPT-5 cung cấp một cái nhìn tổng quan toàn diện về các khả năng và tính năng của AI này. Mặc dù các chi tiết cụ thể về kiến trúc và chức năng của GPT-5 không được cung cấp, system card vẫn nêu bật các ứng dụng và lợi ích tiềm năng của AI. GPT-5 được định vị là một cải tiến đáng kể so với các phiên bản tiền nhiệm, với hiệu suất và các tính năng an toàn được nâng cao. System card nhấn mạnh khả năng của AI trong việc xử lý các tác vụ phức tạp và cung cấp các phản hồi chính xác, nhiều thông tin. Tuy nhiên, bản chất chính xác của những cải tiến này và các tính năng cụ thể giúp chúng hoạt động không được tiết lộ. Việc phát hành GPT-5 đánh dấu một cột mốc quan trọng đối với OpenAI, thể hiện sự đầu tư liên tục của công ty vào nghiên cứu và phát triển AI. Giống như bất kỳ công nghệ AI mới nào, các tác động và ứng dụng tiềm năng của GPT-5 là rất lớn, và nhiều thông tin hơn có thể sẽ được tiết lộ trong những ngày và tuần tới.

4

The Download: Các cuộc tranh luận về nút thắt cổ chai của AI, và các thử nghiệm BCI cất cánh

The Download: AI bottleneck debates, and BCI trials take off

MIT Tech Review Đọc bài viết →

Điểm tin công nghệ tuần này nêu bật một số phát triển quan trọng. Startup AI Subquadratic đã đạt được đột phá trong việc giải quyết một nút thắt cổ chai toán học đã cản trở các large language model trong gần một thập kỷ. Phương pháp của họ giảm số lượng phép tính cần thiết, giúp model nhanh hơn, rẻ hơn và tiết kiệm năng lượng hơn. Mặc

> dù một số chuyên gia vẫn còn hoài nghi, công ty đã bắt đầu chia sẻ nghiên cứu của mình, khơi dậy sự quan tâm trong lĩnh vực này. Trong khi đó, các thử nghiệm brain-computer interface (BCI) đang tăng tốc, với Trung Quốc trở thành quốc gia đầu tiên phê duyệt BCI cho mục đích y tế. Một người đàn ông mắc ALS, Casey Harrell, đã trở thành "power user" của một thiết bị BCI, giúp anh duy trì thu nhập và kết nối lại với những người thân yêu. Công nghệ này đang phát triển nhanh chóng, với các engineer cung cấp nhiều tính năng hơn bao giờ hết. Các câu chuyện đáng chú ý khác bao gồm việc các nhân viên Amazon có thể bị sa thải vì ủng hộ giới hạn data center, và một phát hiện hóa thạch mới đã viết lại 150 năm lý thuyết tiến hóa. Ngoài ra, Thượng nghị sĩ Mỹ Bernie Sanders đã đề xuất luật trao quyền sở hữu trực tiếp các công ty AI cho công chúng thông qua một quỹ AI sovereign wealth fund.

5 Fine-tuning quên kiến thức. RAG rò rỉ context. Hypernetworks xây dựng model mà agent của bạn cần theo yêu cầu.

Fine-tuning forgets. RAG leaks context. Hypernetworks build the model your agent needs on demand.

VentureBeat Đọc bài viết →

> Các team doanh nghiệp thường gặp khó khăn với các AI agent không mang lại hiệu quả như đã hứa. Sau thành công ban đầu, các agent này yêu cầu sự can thiệp của con người để bổ sung context và kiểm tra output, khiến chúng kém hiệu quả hơn. Vấn đề này bắt nguồn từ một vấn đề cơ bản: khi các AI model xử lý nhiều input hơn, độ chính xác của chúng giảm do bản chất của các attention mechanism. Các giải pháp hiện tại, như fine-tuning và in-context learning, đều có những hạn chế. Fine-tuning liên quan đến việc "nướng" kiến thức vào các weight của model, nhưng nó dễ bị catastrophic forgetting, nơi thông tin mới làm xói mòn kiến thức hiện có. In-context learning, cung cấp các policy liên quan tại runtime, dễ bị context rot, nơi các lỗi retrieval không thể phân biệt được với các câu trả lời tự tin. Một phương pháp mới, hypernetworks, tạo ra một model nhỏ, chuyên biệt cho tác vụ theo yêu cầu từ các policy của công ty tại thời điểm inference. Phương pháp này tránh được chi phí retraining của fine-tuning và giới hạn context của prompting. Hypernetworks có tiềm năng khép lại vòng lặp về vấn đề năng lực của agent, cho phép tự chủ và hiệu quả hơn trong các workflow AI. Tuy nhiên, phương pháp này vẫn còn ở giai đoạn đầu, và các câu hỏi về calibration, scale và quyền sở hữu của tài sản đang được cải thiện vẫn còn bỏ ngỏ.

6 The Atlantic đã tạo một database có thể tìm kiếm về âm nhạc dùng để train AI

The Atlantic created a searchable database of the music used to train AI

The Verge AI Đọc bài viết →

> The Atlantic đã tạo một database có thể tìm kiếm về âm nhạc được sử dụng để train các AI model. Phóng viên Alex Reisner của Atlantic đã phát hiện bốn dataset âm nhạc lớn, tổng cộng hơn 21 triệu track, trước đây đã có sẵn miễn phí trực tuyến nhưng không dễ dàng truy cập để sử dụng trong training AI. Các dataset này đã được tải xuống hàng nghìn lần và đã được các công ty như Google và Stability sử dụng trong các bài nghiên cứu. Tuy nhiên, việc sử dụng các dataset này cho mục đích thương mại có thể yêu cầu cấp phép, vì một số nguồn, như Free Music Archive, có các hạn chế. Database bao gồm âm nhạc từ

nhiều nghệ sĩ khác nhau, bao gồm Lady Gaga, Radiohead và Bruce Springsteen. Database hiện đã có sẵn trên trang AI Watchdog của Atlantic, cho phép người dùng tìm kiếm các bài hát và các media khác đang được sử dụng để train các AI model.

7

Vượt ra ngoài LoRA: Bạn có thể đánh bại kỹ thuật fine-tuning phổ biến nhất không?

Beyond LoRA: Can you beat the most popular fine-tuning technique?

Hugging Face Blog Đọc bài viết →

Fine-tuning một open model trên dữ liệu cá nhân có thể tốn nhiều memory, nhưng các kỹ thuật parameter-efficient fine-tuning (PEFT) đã xuất hiện để giảm yêu cầu này. Một kỹ thuật phổ biến, LoRA (Low Rank Adaptation), thêm một vài parameter lên trên base model và chỉ train những parameter đó. Mặc dù được sử dụng rộng rãi, không rõ liệu LoRA có phải là lựa chọn tốt nhất hay không. Nhiều nhà nghiên cứu tuyên bố kỹ thuật của họ vượt trội hơn LoRA, nhưng những tuyên bố này thường bị thiên vị do áp lực phải tạo ra kết quả tốt hơn. Ngoài ra, mỗi nghiên cứu so sánh các bộ kỹ thuật

8

Cách các giới hạn pull request giúp giảm bớt "nhiễu"

How pull request limits are cutting down the noise

GitHub Blog Đọc bài viết →

GitHub đã giới thiệu giới hạn yêu cầu kéo để quản lý khối lượng đóng góp ngày càng tăng vào các kho mã nguồn mở. Với nhiều người đóng góp hơn bao giờ hết, thách thức nằm ở việc phân biệt các đóng góp chất lượng cao với tiếng ồn chất lượng thấp. Tính năng mới này đặt số lượng yêu cầu kéo tối đa mà một người dùng không có quyền ghi có thể có trong một kho tại bất kỳ thời điểm nào. Khi đạt đến giới hạn này, người dùng phải đóng hoặc hợp nhất một yêu cầu kéo trước khi mở một yêu cầu khác. Điều này bao gồm cả các yêu cầu kéo được mở bởi các đại lý AI, chẳng hạn như Copilot. Các nhà đóng góp đáng tin cậy có thể được miễn khỏi các giới hạn này thông qua danh sách bỏ qua. Mục tiêu của giới hạn yêu cầu kéo là giúp quản lý dòng chảy của các đóng góp và làm cho việc xác định các đóng góp có giá trị dễ dàng hơn giữa tiếng ồn.

TIPS & TRICKS CHO DEV

Orchestrating AI Agents

Vấn đề: Xử lý task phức tạp yêu cầu phối hợp nhiều AI agents.
Cách làm: Sử dụng LangGraph để định nghĩa workflow và CrewAI để quản lý agents. Ví dụ, prompt "Design a workflow for sentiment analysis using multiple AI agents" có thể được sử dụng để khởi tạo quá trình.
Đánh giá: Hiệu quả cao khi xử lý task phức tạp, nhưng có thể tăng độ phức tạp của hệ thống.

AutoGen Configuration

Vấn đề: Cấu hình AutoGen cho các task xử lý ngôn ngữ tự nhiên.
Cách làm: Sử dụng lệnh CLI `autogen configure` để thiết lập cấu hình cho AutoGen, sau đó sử dụng `autogen train` để huấn luyện model.
Đắn giá: Hiệu quả khi xử lý task ngôn ngữ tự nhiên, nhưng cần phải có kiến thức về cấu hình và huấn luyện model.

phiData Integration

Vấn đề: Tích hợp phiData vào hệ thống hiện có để tăng cường khả năng xử lý dữ liệu.
Cách làm: Sử dụng API `phiData.connect()` để kết nối với hệ thống và `phiData.process()` để xử lý dữ liệu.
Đánh giá: Hiệu quả khi cần xử lý dữ liệu lớn, nhưng cần phải có kiến thức về API và hệ thống hiện có.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tích hợp AI API vào ứng dụng**
Dev cần biết cách tích hợp AI API vào ứng dụng để tận dụng khả năng của trí tuệ nhân tạo trong việc tự động hóa và cải thiện hiệu suất. Điều này giúp phát triển ứng dụng thông minh hơn và đáp ứng nhu cầu người dùng tốt hơn.

Ví dụ, bạn có thể sử dụng API của LangGraph để tích hợp chức năng tự động hoàn thiện code vào ứng dụng phát triển của mình.

3. Ví dụ thực tế: bạn có thể tạo một ứng dụng hỗ trợ viết code bằng cách tích hợp API của GitHub Copilot, cho phép người dùng nhận đề xuất code khi nhập.

4. Tip hoặc bước tiếp theo: Để bắt đầu tích hợp AI API, hãy chọn một công cụ phù hợp với nhu cầu của ứng dụng, như GitHub Copilot hoặc Claude Code, và tham khảo tài liệu API để thực hiện tích hợp thành công.

Nguồn: Google News · Groq AI