



Sự xuất hiện của DeepSeek đặt ra thách thức lớn đối với các mô hình AI hiện có trên toàn cầu.

XU HƯỚNG PHÁT TRIỂN TRÍ TUỆ NHÂN TẠO VỚI CHI PHÍ THẤP

Phạm Hồng Sơn

Đại học Quốc gia TP Hồ Chí Minh

“

Tháng 01/2025, DeepSeek đã ra mắt ứng dụng chatbot miễn phí đầu tiên dựa trên mô hình DeepSeek-V3, đánh dấu một bước ngoặt quan trọng trong lĩnh vực trí tuệ nhân tạo (AI) toàn cầu. Sự xuất hiện của DeepSeek không chỉ tạo ra những tác động đáng kể mà còn đặt ra thách thức lớn đối với các mô hình AI của Mỹ, đặc biệt về chi phí sử dụng và khả năng tiếp cận người dùng. Trong bối cảnh này, khái niệm “AI với chi phí thấp” đã trở thành một chủ đề nóng, với các mô hình được vận hành với chi phí thấp thông qua tối ưu hóa thuật toán, sử dụng phần cứng tiết kiệm chi phí và các chiến lược giảm tải tài nguyên.

”

Bước ngoặt cạnh tranh

DeepSeek-V3 và các phiên bản liên quan của nó đã được đánh giá ngang hàng với những chatbot hàng đầu đến từ các tập đoàn công nghệ Mỹ như OpenAI hay Google. Điều đáng chú ý là mô hình này chỉ tiêu tốn khoảng 6 triệu USD cho quá trình phát triển, minh chứng cho khả năng đạt hiệu suất ấn tượng mà không cần đầu tư quá lớn. Thành công này đã khơi mào cho cuộc chạy đua toàn cầu trong lĩnh vực phát triển AI chi phí thấp, thu hút sự tham gia của hàng loạt doanh nghiệp và tổ chức công nghệ.

Sự xuất hiện của DeepSeek đã thúc đẩy các gã khổng lồ như OpenAI và Google gia nhập thị trường AI với chi phí thấp. Đáng chú ý, OpenAI đã trình làng o3-mini vào ngày 03/02/2025 - phiên bản được tối ưu hóa với mức phí vận hành chỉ bằng 50% so với phiên bản tiền nhiệm o1-mini. Không dừng lại ở việc cắt giảm chi phí, o3-mini còn sở hữu khả năng xử lý tác vụ phức tạp vượt trội nhờ tốc độ phản hồi nhanh và độ trễ thấp, đánh dấu bước đột phá về hiệu quả tính toán trong bối cảnh DeepSeek đang lên ngôi. Điểm nhấn ấn tượng của o3-mini nằm ở cơ chế điều chỉnh nỗ lực suy luận, cho phép người dùng tùy biến mức độ xử lý từ thấp đến cao dựa trên yêu cầu nhiệm vụ. Tính năng này mở ra khả năng ứng dụng đa dạng, từ phân tích dữ liệu cơ bản đến hỗ trợ ra quyết định cấp chiến lược. Sự cạnh tranh gay gắt giữa các nền tảng AI không chỉ thúc đẩy làn sóng đổi mới công nghệ mà còn thiết lập xu hướng phát triển bền vững, nơi hiệu suất và tính kinh tế song hành.

Không chịu kém cạnh, Google đã chính thức trình làng bộ đôi Gemini 2.0 Flash và Flash-lite vào ngày 5/2/2025. Phiên bản Flash gây ấn tượng với hiệu suất cao cùng độ trễ cực thấp, tích hợp khả năng xử lý đa phương thức (hình ảnh, video, âm thanh) và tạo đầu ra video đa định dạng tương ứng. Phiên bản Flash-lite được tối ưu hóa với mức giá chỉ bằng 50% bản tiêu chuẩn, nhưng vẫn giữ nguyên ưu điểm xử lý đa nhiệm và tốc độ phản hồi. Điều này biến nó thành giải pháp chatbot AI có chi phí lý tưởng cho doanh nghiệp cần cân bằng giữa hiệu năng và ngân sách, đặc biệt phù hợp với các ứng dụng thời gian thực như xử lý video trực tiếp hay phân tích dữ liệu cảm biến. Trong khi o3-mini của OpenAI tỏa sáng với thể mạnh xử lý ngôn ngữ tự nhiên và ứng dụng STEM chuyên sâu, bộ đôi Gemini của Google lại chiếm lợi thế nhờ kiến trúc đa phương thức linh hoạt và khả năng tối ưu hóa chi phí vượt trội. Sự phân hóa này không chỉ phản ánh xu hướng phát triển đa dạng của thị trường AI với chi phí thấp mà còn mở ra nhiều lựa chọn tối ưu cho người dùng tùy theo nhu cầu ứng dụng.



Các gã khổng lồ như OpenAI và Google bắt đầu gia nhập thị trường mô hình AI có chi phí thấp.

Định hướng phát triển mô hình AI với chi phí thấp

Sự cạnh tranh giữa các mô hình AI với chi phí thấp không chỉ thúc đẩy đổi mới công nghệ mà còn mở rộng cơ hội ứng dụng AI trên quy mô toàn cầu. Những mô hình này đang định hình một cuộc cách mạng AI với chi phí thấp chưa từng có, giúp doanh nghiệp và cá nhân dễ dàng triển khai AI mà không cần đầu tư lớn vào hạ tầng hoặc nghiên cứu chuyên sâu.

Trong tương lai gần, chắc chắn cuộc đua phát triển AI với chi phí thấp sẽ tiếp tục diễn ra sôi nổi. Các công ty công nghệ phải liên tục đổi mới và cải thiện hiệu suất của các mô hình AI trong khi vẫn giữ chi phí ở mức hợp lý. Điều này cho thấy, AI không còn là một công nghệ đắt đỏ, chỉ dành cho các tổ chức lớn. Mô hình DeepSeek là một ví dụ điển hình cho sự phát triển của AI với chi phí thấp. Mô hình này đã chứng minh rằng AI có thể được tạo ra với chi phí thấp mà không phải hy sinh hiệu suất. Ngoài ra, các mô hình như OpenAI o3-mini và Google Gemini 2.0 Flash cũng đã được nghiên cứu và phát triển, mở ra nhiều cơ hội ứng dụng trong các lĩnh vực khác nhau của cuộc sống.

Sự phát triển của các mô hình AI với chi phí thấp cũng đang tạo ra nhiều cơ hội cho các nhà nghiên cứu và phát triển. Tại Mỹ, Đại học Stanford và Washington đã phát triển mô hình "s1" với chi phí chỉ khoảng 50 USD. Một nhóm nghiên cứu tại Đại học California, Berkeley cũng đã tái tạo thành công công nghệ cốt lõi đằng sau mô hình AI của DeepSeek, đặt tên là "Tiny Zero", có tổng chi phí nghiên cứu chỉ khoảng 30 USD.

DeepSeek là một trong số ít các mô hình AI mã nguồn mở miễn phí hoặc có chi phí thấp nhưng vẫn đảm bảo hiệu suất cao. Điều này cho phép cộng đồng tiếp cận mã nguồn miễn phí và tùy chỉnh theo nhu cầu riêng, giúp giảm chi phí triển khai AI cho doanh nghiệp. Đây là một sự khác biệt so với các mô hình AI với chi phí thấp khác như OpenAI o3-mini và Google Gemini Flash, vốn không công khai mã nguồn. Sự cạnh tranh giữa các mô hình AI với chi phí thấp không

chỉ thúc đẩy đổi mới công nghệ, mà còn mở ra nhiều cơ hội ứng dụng trên toàn cầu. Tuy nhiên, vẫn còn những thách thức về chi phí vận hành thực tế, vì nhiều mô hình AI với chi phí thấp vẫn đang tận dụng nguồn lực sẵn có để giảm chi phí.

Có thể thấy, cuộc đua trong việc phát triển AI với chi phí thấp đang diễn ra sôi nổi. Mặc dù vẫn còn những thách thức về chi phí vận hành thực tế, nhưng các mô hình AI với chi phí thấp đã mở ra nhiều cơ hội ứng dụng trên toàn cầu, đặt nền móng cho một cuộc cách mạng AI với chi phí tiếp cận thấp hơn bao giờ hết. Trong tương lai, sự cạnh tranh và đổi mới trong lĩnh vực này sẽ tiếp tục diễn ra, mang lại nhiều lợi ích cho cả doanh nghiệp và người dùng.

*

* *

Mặc dù sự cạnh tranh trong nghiên cứu AI với chi phí thấp đang thúc đẩy đổi mới, nó cũng đặt ra những thách thức quan trọng cần giải quyết. Các vấn đề đạo đức, hệ thống quản lý và tính bền vững ngày càng trở thành chủ đề mang lại nhiều tranh luận, đặc biệt khi AI có tầm ảnh hưởng sâu rộng hơn. Việc phát triển AI một cách có trách nhiệm là yếu tố quan trọng để duy trì niềm tin và tránh những hậu quả không mong muốn. Trong tương lai, nghiên cứu AI có thể sẽ chứng kiến sự hợp tác ngày càng nhiều hơn giữa các nhà phát triển, doanh nghiệp và các nhà hoạch định chính sách. Tương lai của AI sẽ được định hình bởi những bước tiến trong tự động hóa nghiên cứu, mở rộng khả năng tiếp cận và cam kết về phát triển AI có trách nhiệm - đặc biệt trong các lĩnh vực tài chính, blockchain và công nghệ.