BIS - MT · 17/08/2026

# Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

*“It does not matter how slowly you go as long as you do not stop.”*

↪ Không quan trọng bạn đi chậm đến đâu, miễn là bạn không dừng lại.

— Khổng Tử

*Tốc độ không quan trọng bằng sự kiên trì — người đi chậm mà bền vững sẽ đến đích trước người đi nhanh mà bỏ cuộc.*

## TIN TỨC NỔI BẬT

1

### Hướng Dẫn Chạy Local AI Models trên RTX 5060 Ti của Bạn — Từng Bước Một

*How to Run Local AI Models on Your RTX 5060 Ti — Step-by-Step Guide*

TechnoSports Media Group  Đọc bài viết →

Bài viết "Hướng dẫn từng bước chạy mô hình AI cục bộ trên RTX 5060 Ti của bạn" cung cấp hướng dẫn toàn diện cho người dùng để chạy mô hình AI cục bộ trên card đồ họa NVIDIA RTX 5060 Ti của họ. Để bắt đầu, người dùng cần cài đặt phần mềm cần thiết, bao gồm CUDA, cuDNN và Trình điều khiển NVIDIA. Bài viết sau đó phác thảo quá trình thiết lập môi trường, bao gồm cài đặt các gói cần thiết và cấu hình hệ thống. Tiếp theo, người dùng được hướng dẫn qua quá trình cài đặt và thiết lập các Framework Học sâu như TensorFlow hoặc PyTorch, những thứ cần thiết để chạy mô hình AI. Bài viết cũng đề cập đến cách biên dịch và chạy mô hình AI, bao gồm cung cấp ví dụ mã và mẹo để tối ưu hóa hiệu suất. Trong suốt hướng dẫn, bài viết nhấn mạnh tầm quan trọng của việc đảm bảo rằng hệ thống được cấu hình đúng và đáp ứng các yêu cầu phần cứng và phần mềm cần thiết. Bằng cách làm theo hướng dẫn từng bước, người dùng có thể chạy thành công mô hình AI cục bộ trên card đồ họa RTX 5060 Ti của họ.

2

### Command Layer AI Đa Agent cho Kho Hàng: Nâng Tầm Hiệu Suất Vận Hành 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 phát triển một Lớp Lệnh AI Kho Hàng Nhiều Chủ Thể, một công nghệ tiên tiến được thiết kế để nâng cao hiệu quả hoạt động và trí tuệ chuỗi cung ứng trong các kho hàng. Giải pháp đổi mới này tận dụng trí tuệ nhân tạo (AI) để tối ưu hóa hoạt động kho hàng, tự động hóa quy trình và cải thiện hiệu suất tổng thể. Lớp Lệnh AI Kho Hàng Nhiều Chủ Thể là một lớp lệnh tích hợp nhiều tác nhân AI để quản lý các nhiệm vụ kho hàng khác nhau, bao gồm quản lý hàng tồn kho, thực hiện đơn hàng và hậu cần. Bằng cách tự động hóa các nhiệm vụ này, hệ thống giảm thiểu sai sót thủ công, tăng năng suất và nâng cao khả năng ra quyết định. Lợi ích chính của Lớp Lệnh AI Kho Hàng Nhiều Chủ Thể bao gồm: - Cải thiện hiệu quả hoạt động thông qua quản lý nhiệm vụ tự động - Nâng cao trí tuệ chuỗi cung ứng thông qua phân tích dữ liệu thời gian thực - Tăng năng suất và giảm sai sót thủ công - Cải thiện khả năng ra quyết định thông qua thông tin dựa trên dữ liệu Lớp Lệnh AI Kho Hàng Nhiều Chủ Thể có tiềm năng cách mạng hóa hoạt động kho hàng, cho phép các doanh nghiệp đạt được sự xuất sắc hoạt động và dẫn đầu trong thị trường cạnh tranh.

3

## Giới Thiệu Open Agent Specification (Agent Spec): Một Đại Diện Thống Nhất cho AI Agents

*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 nhằm tiêu chuẩn hóa cách thiết kế, phát triển và tích hợp các tác nhân AI vào các hệ thống khác nhau. Thông số kỹ thuật này cung cấp một khuôn khổ chung để mô tả hành vi, khả năng và tương tác của các tác nhân AI, cho phép giao tiếp và cộng tác liền mạch giữa các tác nhân và hệ thống khác nhau. Agent Spec được thiết kế để linh hoạt và thích ứng, cho phép nó hỗ trợ các loại tác nhân AI khác nhau, bao gồm cả những tác nhân dựa trên học máy, hệ thống dựa trên quy tắc và các phương pháp tiếp cận kết hợp. Bằng cách cung cấp một biểu diễn thống nhất, Agent Spec tạo điều kiện cho việc tạo ra các hệ thống AI tinh vi và hiệu quả hơn, có thể tương tác với nhau và với con người một cách trực quan và hiệu quả hơn. Việc giới thiệu Agent Spec dự kiến sẽ đẩy nhanh sự phát triển và triển khai các ứng dụng AI, cho phép các tổ chức tận dụng tối đa tiềm năng của công nghệ AI và học máy.

4

## AI agent là gì?

*What is an AI agent?*

LangChain Blog  Đọc bài viết →

Trong lĩnh vực trí tuệ nhân tạo, định nghĩa của một tác nhân AI có thể thay đổi tùy thuộc vào ngữ cảnh. Tuy nhiên, một định nghĩa thực tế là một tác nhân AI là một hệ thống nơi mô hình có quyền kiểm soát luồng điều khiển, xác định các bước và hành động tiếp theo của chính nó. Điều này có thể bao gồm từ một bộ định tuyến mô hình ngôn ngữ đơn giản đến một hệ thống chọn công cụ của riêng nó và chạy trong khoảng thời gian dài. Mức độ tự chủ trong một tác nhân AI phụ thuộc vào mức độ kiểm soát mà mô hình có đối với đầu ra, bước tiếp theo và công cụ có sẵn. Một phổ tự chủ đã được đề xuất, với sáu mức độ, bao gồm các quy trình làm việc, máy trạng thái và các tác nhân tự chủ hoàn toàn. Các phương tiện tự hành có một hệ thống phân cấp tương tự, với các mức độ của hệ thống hỗ trợ lái xe, nhưng ngành công nghiệp AI vẫn chưa thống nhất về một hệ thống tiêu chuẩn hóa. Trong sản xuất, sự khác biệt giữa một tác nhân AI và một quy trình làm việc là một quyết định thiết kế. Một nhóm nên sử dụng một tác nhân khi vấn đề yêu cầu mô hình phải giải thích đầu vào không có cấu trúc hoặc chọn hành động tiếp theo, và sử dụng mã quyết định cho các bước có yêu cầu rõ ràng. Một tác nhân AI bao gồm bốn thành phần: mô hình, công cụ, bộ nhớ và vòng lặp suy luận, làm việc cùng nhau để cho phép mô hình kiểm soát luồng.

5

## Computer History của ChatGPT theo dõi các cú click và keystrokes của bạn

*ChatGPT's Computer History tracks your clicks and keystrokes*

The Verge AI  Đọc bài viết →

ChatGPT, một trợ lý AI phổ biến, đã giới thiệu một tính năng mới gọi là Lịch sử Máy tính trong ứng dụng máy tính để bàn trên macOS. Tính năng này theo dõi các lần nhấp và keystrokes của người dùng, học hỏi thói quen làm việc của họ và đề xuất các tự động hóa để cải thiện năng suất. Nó tạo ra một dòng thời gian của các hoạt động của người dùng, mà ChatGPT và người bạn đồng hành AI của nó, Codex, có thể tham khảo khi thực hiện yêu cầu. Người dùng có quyền kiểm soát tính năng này, cho phép họ chọn tham gia, loại trừ các ứng dụng và trang web nhất định, và xóa các mục để có quyền kiểm soát tốt hơn. OpenAI, công ty đứng sau ChatGPT, đã đảm bảo rằng Lịch sử Máy tính

sẽ không ghi lại hình ảnh, video hoặc âm thanh, thay vào đó dựa trên "sự kiện". Tính năng này gợi nhớ đến Windows Recall của Microsoft, nhưng không sử dụng ảnh chụp màn hình. Tính năng này nhằm cung cấp trải nghiệm được cá nhân hóa hơn cho người dùng, nhưng một số người có thể thấy nó "bất an".

6

## Unsloth Biến Mẹo Fine-Tuning Thành Desktop App Chạy Model 744B-Parameter trên Một GPU

*Unsloth Turned a Fine-Tuning Trick Into a Desktop App That Runs 744B-Parameter Models on One GPU*

Dev.to AI Đọc bài viết →

Dự án Unsloth, trước đây ít được biết đến, đã trải qua một sự thay đổi đáng kể. Dự án ban đầu tập trung vào một thư viện Python để tinh chỉnh các mô hình ngôn ngữ lớn (LLM) trên một GPU duy nhất, đã phát triển thành một ứng dụng máy tính để bàn đa nền tảng. Ứng dụng Unsloth Desktop cho phép người dùng chạy, đào tạo và cung cấp các mô hình lớn cục bộ, bao gồm cả mô hình có 744 tỷ tham số. Ứng dụng được xây dựng trên một framework gốc dựa trên Tauri và không yêu cầu môi trường Python hoặc thiết lập thủ công. Nó có giao diện trò chuyện, trình duyệt mô hình và đường ống đào tạo, khiến nó trở thành một công cụ toàn diện cho phát triển AI cục bộ. Unsloth hiện bao gồm ba sản phẩm: Unsloth Core, một thư viện Python để tinh chỉnh và đào tạo nhanh; Unsloth Studio, một giao diện web tự tổ chức để lập danh mục mô hình, công cụ dữ liệu và xuất; và Unsloth Desktop, một ứng dụng gốc bao gồm Studio trong một giao diện thân thiện với người dùng. Phương pháp GGUF động của dự án cho phép tải mô hình lớn trên phần cứng tiêu dùng hoặc prosumer, bảo tồn độ chính xác ở tốc độ bit thấp. Unsloth hỗ trợ nhiều nền tảng phần cứng, bao gồm NVIDIA, AMD và Apple Silicon, và cung cấp nhiều tùy chọn đào tạo, bao gồm đào tạo LoRA, QLoRA và FP8.

7

## Sau Orthogonality: Virtue-Ethical Agency và AI Alignment

*After Orthogonality: Virtue-Ethical Agency and AI Alignment*

The Gradient Đọc bài viết →

Bài viết "Sau Tính Chính Orthogonal: Cơ quan Đạo đức - Đức hạnh và Sự Liên kết AI" cho rằng con người hợp lý và AI không nên có mục tiêu, mà thay vào đó nên liên kết hành động của họ với các thực hành. Những thực hành này là mạng lưới các hành động, xu hướng hành

động, tiêu chí đánh giá và tài nguyên cấu trúc và thúc đẩy bản thân. Tác giả đề xuất rằng sự suy xét của các tác nhân AI nên chia sẻ một logic tương tự với các thực hành của con người để thực sự hỗ trợ và cộng tác với cơ quan con người. Cách tiếp cận này rất quan trọng để liên kết AI với các thuộc tính an toàn như minh bạch, hữu ích và vô hại. Bài viết cũng khám phá khái niệm eudaimonia, hay sự thịnh vượng con người tích cực, hợp lý và ý nghĩa của nó đối với sự liên kết AI. Tác giả cho rằng eudaimonia không phải là một trạng thái hoặc quỹ đạo mong muốn, mà thay vào đó là một cấu trúc suy xét khác với tính hợp lý kết quả tiêu chuẩn. Hình thức hoạt động hợp lý này, được gọi là tính hợp lý eudaimonic, được coi là một khuôn khổ hữu ích cho AI được liên kết với con người, mang lại lợi thế ổn định và an toàn so với các hình thức cơ quan khác. Tác giả tuyên bố rằng tính hợp lý eudaimonic là một hình thức cơ quan tự nhiên, hiệu quả ngay cả theo tiêu chuẩn kết quả, và việc liên kết AI với hình thức tính hợp lý này có thể giải quyết các vấn đề an toàn AI cổ điển và nghịch lý. Bài viết kết thúc bằng cách nhấn mạnh tầm quan trọng của việc xem xét tính tự nhiên của cơ quan và tính hợp lý trong sự liên kết AI.

8

## Ghi lại, train và deploy từ một nơi với Strands Agents, LeRobot và Hugging Face Storage Buckets

*Record, train, and deploy from one place with Strands Agents, LeRobot, and Hugging Face Storage Buckets*

Hugging Face Blog  Đọc bài viết →

> AWS đã phát triển một SDK mã nguồn mở có tên là Strands Robots, cho phép người dùng tạo thành các trừu tượng hóa robot, mô phỏng và ngăn xếp LeRobot thành một tác nhân duy nhất. Tác nhân này có thể ghi lại các bản demo, đào tạo các chính sách và triển khai chúng lên các robot vật lý. Một tính năng mới, Strands Agents, cho phép người dùng ghi lại các bản demo vào một Hugging Face Storage Bucket, sử dụng deduplication ở mức byte để giảm thiểu chi phí truyền dữ liệu. Tác nhân sau đó có thể đào tạo trên tập dữ liệu bằng cách truyền trực tiếp từ Hugging Face Hub, thay vì tải nó xuống. Một khi được đào tạo, tác nhân có thể triển khai chính sách trở lại robot với các thay đổi tối thiểu. Quá trình này có thể được lặp lại liên tục, cho phép người dùng thu thập các tập dữ liệu, đào tạo các chính sách và triển khai chúng trong một vòng lặp. Tác nhân cũng có thể lưu trữ tập dữ liệu trong cùng định dạng LeRobot trên đĩa trong suốt quá trình, giúp việc quản lý và tái sử dụng dữ liệu trở nên dễ dàng hơn. Cách tiếp cận này có thể giúp giảm chi phí truyền dữ liệu và cải thiện hiệu quả của quá trình đào tạo.

---

TIPS & TRICKS CHO DEV

### Quản lý context window

**Vấn đề:** Context window quá nhỏ khiến mô hình AI không thể nắm bắt thông tin cần thiết.

**Cách làm:** Sử dụng kỹ thuật context window management, ví dụ như tăng kích thước context window khi sử dụng mô hình ngôn ngữ transformer.

**Đánh giá:** Hiệu quả khi cần xử lý văn bản dài, nhưng có thể tăng chi phí tính toán.

### Tối ưu hóa long-context

**Vấn đề:** Long-context quá dài gây ra sự lãng phí tài nguyên và giảm hiệu suất.

**Cách làm:** Sử dụng kỹ thuật cắt đoạn văn bản và chỉ giữ lại thông tin quan trọng, ví dụ như sử dụng lệnh CLI `--max-context-length`.

**Đánh giá:** Hiệu quả khi cần giảm độ phức tạp, nhưng có thể mất thông tin quan trọng.

### Sử dụng memory hiệu quả

**Vấn đề:** Bộ nhớ không đủ để lưu trữ thông tin context.

**Cách làm:** Sử dụng kỹ thuật lưu trữ bộ nhớ ngoài, ví dụ như sử dụng `disk-based caching` khi sử dụng mô hình AI như Claude.

**Đánh giá:** Hiệu quả khi cần xử lý dữ liệu lớn, nhưng có thể giảm tốc độ truy xuất.

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để xây dựng ứng dụng AI hiệu quả, các nhà phát triển cần biết cách tối ưu hóa chi phí và hiệu năng của Large Language Model (LLM). Điều này giúp giảm thiểu chi phí vận hành và cải thiện trải nghiệm người dùng. Việc tối ưu hóa LLM cũng giúp các ứng dụng AI hoạt động mượt mà hơn trên các thiết bị có tài nguyên hạn chế.
3. Ví dụ, chúng ta có thể sử dụng kỹ thuật fine-tuning và LoRA để tối ưu hóa LLM cho các use case cụ thể, giúp giảm thiểu kích thước mô hình và cải thiện hiệu suất.
4. Tip hoặc bước tiếp theo: Hãy bắt đầu bằng cách đánh giá các mô hình LLM hiện có và xác định các khu vực cần tối ưu hóa, sau đó áp dụng các kỹ thuật như fine-tuning, pruning và quantization để cải thiện hiệu năng và giảm chi phí.