BIS - MT · 05/08/2026

# Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

*“Đã trót nhúng chàm thì phải giặt cho sạch.”*

— Tục ngữ Việt Nam

*Khi đã mắc sai lầm, hãy dũng cảm nhận trách nhiệm và sửa chữa — đó là biểu hiện của nhân cách trưởng thành.*

## TIN TỨC NỔI BẬT

1

### Cách chạy các AI model cục bộ trên RTX 5060 Ti của bạn — Hướng dẫn từng bước

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

technosports.co.in Đọc bài viết →

> Không có nội dung được cung cấp. Vui lòng cung cấp nội dung cho bài viết "Hướng dẫn từng bước chạy mô hình AI cục bộ trên RTX 5060 Ti" để tôi có thể viết một bản tóm tắt rõ ràng và đầy đủ thông tin.

2

### Giới thiệu Open Agent Specification (Agent Spec): Một chuẩn thống nhất cho AI agent

*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 cung cấp một khuôn khổ tiêu chuẩn hóa để định nghĩa và tương tác với các tác nhân AI trên nhiều nền tảng và ứng dụng khác nhau. Thông số kỹ thuật này được thiết kế để mã nguồn mở và do cộng đồng dẫn dắt, cho phép các nhà phát triển đóng góp và định hình tương lai của việc phát triển tác nhân AI. Agent Spec được xây dựng trên các công nghệ hiện có như JSON và Protocol Buffers, giúp dễ dàng tích hợp với các hệ thống hiện có. Nó cung cấp một tập hợp toàn diện các định nghĩa và API để tạo, quản lý và tương tác với các tác nhân AI. Điều này bao gồm các tính năng như danh tính tác nhân, khả năng và hành vi, cũng như API cho các nhiệm vụ như tạo, cập nhật và xóa tác nhân. Việc giới thiệu Agent

Spec dự kiến sẽ thúc đẩy sự phát triển của các tác nhân AI tinh vi và tương tác hơn, cho phép tích hợp liền mạch trên các nền tảng và ứng dụng khác nhau. Bằng cách cung cấp một biểu diễn thống nhất cho các tác nhân AI, Agent Spec có tiềm năng tăng tốc đổi mới trong lĩnh vực trí tuệ nhân tạo.

## 3 Agent Factory: Kết nối các agent, app và data với 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 nền tảng được thiết kế để kết nối các tác nhân, ứng dụng và dữ liệu bằng cách sử dụng các tiêu chuẩn mở mới. Nền tảng này tận dụng các tiêu chuẩn Cloud PC (MCP) và Application to Application (A2A) của Microsoft để tạo điều kiện cho sự tương tác liền mạch giữa các thành phần khác nhau. Agent Factory cho phép các nhà phát triển tạo, triển khai và quản lý các tác nhân có thể tương tác với các ứng dụng và nguồn dữ liệu khác nhau. Sự tích hợp này cho phép trao đổi dữ liệu hiệu quả hơn, quy trình làm việc được tối ưu hóa và khả năng tự động hóa được nâng cao. Việc sử dụng các tiêu chuẩn MCP và A2A đảm bảo rằng các kết nối giữa các tác nhân, ứng dụng và dữ liệu là bảo mật, có thể mở rộng và tương lai. Cách tiếp cận này cũng thúc đẩy khả năng tương tác, cho phép các hệ thống khác nhau giao tiếp hiệu quả và giảm nhu cầu tích hợp tùy chỉnh. Bằng cách tận dụng Agent Factory và các tiêu chuẩn mở liên quan, các nhà phát triển có thể xây dựng các hệ thống mạnh mẽ, linh hoạt và kết nối hơn. Nền tảng này có tiềm năng cách mạng hóa cách thức các ứng dụng và dữ liệu tương tác, cho phép các doanh nghiệp trở nên linh hoạt và phản ứng nhanh hơn với các điều kiện thị trường thay đổi, đặc biệt là khi tích hợp với các công nghệ như AI, API, LLM, model, token, framework, v.v.

## 4 Phiên bản LLM mới bổ sung hỗ trợ reasoning trace, OpenAI Responses, server-side tool và logging thông minh hơn

*New release of LLM adds support for reasoning traces, OpenAI Responses, server-side tools, and smarter logging*

Simon Willison Đọc bài viết →

Một bản cập nhật lớn đã được phát hành cho dự án Large Language Model (LLM), phiên bản 0.32. Bản phát hành này mang lại những cải tiến đáng kể, bao gồm hỗ trợ cho các dấu vết lý luận có thể nhìn thấy, các công cụ phía máy chủ và ghi nhật ký thông minh hơn. Tính năng dấu vết lý luận cho phép người dùng xem quá trình suy nghĩ đằng sau các phản hồi của mô hình, có thể được bật hoặc tắt. Bản cập nhật cũng giới thiệu hỗ trợ cho API OpenAI Responses, cho phép các tính năng và mô hình mới. Gia đình mô hình GPT-5.6 hiện được bao gồm sẵn, và mô hình mặc định được sử dụng với lệnh "prompt" là GPT-5.6 Luna. Các công cụ phía máy chủ từ các nhà cung cấp khác nhau, chẳng hạn như OpenAI, hiện có thể được sử dụng với các cuộc gọi LLM. Bản phát hành bao gồm một số bản cập nhật plugin, bao gồm plugin llm-anthropic, llm-gemini và llm-openrouter. Lệnh llm openai endpoint cho phép người dùng thực thi các lệnh prompt đối với bất kỳ điểm cuối tương thích OpenAI nào dưới dạng một dòng lệnh. Bản cập nhật cũng giới thiệu một tham số model.prompt(messages=[]), có thể được sử dụng để gửi tin nhắn đến mô hình một cách hiệu quả hơn. Ngoài ra, bản phát hành bao gồm một cửa hàng tin nhắn có thể định địa chỉ nội dung mới, được mô hình hóa theo Git, cải thiện việc ghi nhật ký và giảm đầu ra JSON trùng lặp. Bản cập nhật này được coi là đáng kể nhất kể từ khi ra mắt dự án ban đầu, và các plugin hiện có nên tiếp tục hoạt động, nhưng có thể yêu cầu nâng cấp để tham gia đầy đủ vào hệ thống sự kiện truyền trực tuyến mới.

5

## The Download: Hạn chế robot của Mỹ và việc ICE thu thập DNA

*The Download: US robot restrictions, and ICE's DNA grab*

MIT Tech Review Đọc bài viết →

Chính phủ Mỹ đã thực hiện các bước quan trọng trong ngành công nghệ. Ủy ban Thương mại Liên bang đã ban hành lệnh cấm nhập khẩu rô-bốt tiên tiến từ nước ngoài, bao gồm rô-bốt hình người, rô-bốt bốn chân và rô-bốt bánh xe. Động thái này được coi là một biện pháp bảo vệ ngành rô-bốt đang phát triển, vẫn còn trong giai đoạn đầu. Lệnh cấm này không nằm trong chiến lược thương mại truyền thống của Trung Quốc, mà là sự mở rộng của chính quyền Trump trong việc bảo vệ ngành công nghiệp AI. Trong khi đó, Cơ quan Di trú và Hải quan Mỹ (ICE) đã thu thập gần một triệu mẫu DNA của người dân vào năm ngoái, với hầu hết trong số họ không có tiền án. Điều này đã gây ra lo ngại về việc sử dụng dữ liệu DNA và ảnh hưởng của nó đến quyền

riêng tư cá nhân. Trong các tin tức công nghệ khác, Trung Quốc đang có khả năng trở thành một nhà lãnh đạo trong lĩnh vực phát triển phần mềm, ngoài sự thống trị hiện có trong lĩnh vực phần cứng. Các nhà lãnh đạo công nghệ và chính trị Mỹ đang phản ứng với sự phát triển này với sự kết hợp giữa lo ngại và bất ổn. Ngoài ra, việc sử dụng AI đang trở nên phổ biến hơn trong các ngành công nghiệp khác nhau, bao gồm kinh doanh, chăm sóc sức khỏe và vận tải. "Tokenomics" AI là một lĩnh vực mới nhằm đo lường giá trị của các khoản đầu tư AI, trong khi nền kinh tế Mỹ đang trở nên phụ thuộc nhiều hơn vào sự bùng nổ của AI.

6

## Mảng datacenter của AMD bùng nổ trong khi gaming bị xếp sau

*AMD’s datacenter business is booming while gaming takes a backseat*

The Verge AI Đọc bài viết →

AMD đã báo cáo sự tăng trưởng đáng kể trong doanh thu trung tâm dữ liệu của mình, tăng hơn gấp đôi lên 6,7 tỷ đô la trong báo cáo thu nhập mới nhất. Sự tăng trưởng này được thúc đẩy bởi nhu cầu ngày càng tăng về khả năng AI, với công ty dự kiến doanh thu trung tâm dữ liệu sẽ tăng hơn gấp đôi nữa vào năm 2027. Ngược lại, doanh thu trò chơi của AMD đã giảm 31% xuống 779 triệu đô la, do giá tăng và thiếu linh kiện ảnh hưởng đến Xbox Series X/S, PS5 và Steam Deck. Doanh thu tổng thể của công ty đã đạt mức kỷ lục 11,5 tỷ đô la, với kinh doanh trung tâm dữ liệu chiếm 58% doanh thu của công ty trong quý. Doanh thu kinh doanh PC và trò chơi của AMD tăng 6% so với năm ngoái, với doanh thu khách hàng tăng 23% nhờ vào doanh số bán bộ xử lý Ryzen. Giám đốc điều hành của công ty, Lisa Su, cho rằng sự tăng trưởng này là do AI thúc đẩy nhu cầu tính toán trên tất cả các thị trường, đặt AMD vào vị trí tốt để tận dụng cơ hội này và mang lại tăng trưởng doanh thu và lợi nhuận đáng kể trong tương lai.

7

## Vì sao lãng phí R&D vẫn tồn tại dù AI đã được áp dụng rộng rãi?

*Why R&D Waste Persists Despite Widespread AI Adoption*

IEEE Spectrum Đọc bài viết →

Một báo cáo mới, Báo cáo Tiêu chuẩn Nghiên cứu và Phát triển 2026, nhấn mạnh vấn đề lãng phí nghiên cứu và phát triển (R&D) đang diễn

ra dù sự áp dụng trí tuệ nhân tạo (AI) trong các tổ chức ngày càng tăng. Báo cáo, dựa trên một cuộc khảo sát hơn 200 chuyên gia R&D cấp cao, cho thấy nhiều công ty vẫn tiếp tục mất ngân sách đáng kể cho các dự án không bao giờ đạt được thương mại hóa. Mặc dù sử dụng các công cụ AI, các đội đổi mới gặp khó khăn với các nguồn dữ liệu phân mảnh, các quy trình phê duyệt dài dòng và các thông tin quan trọng bị trì hoãn, điều này cản trở khả năng đưa ra quyết định đầu tư thông minh của họ. Kết quả là, các đội thường ưu tiên các cơ hội sai, không loại bỏ các dự án có giá trị thấp sớm và gặp phải những thất bại tốn kém ở giai đoạn cuối. Báo cáo nhằm cung cấp thông tin thực tế và dữ liệu tiêu chuẩn để giúp các tổ chức giảm lãng phí R&D và tăng tốc thời gian đưa sản phẩm ra thị trường.

8

## Canary token và digital tripwire

*Canary tokens and digital tripwires*

Changelog  Đọc bài viết →

Người sáng lập Thinkst, Haroon Meer, thảo luận về các sản phẩm của công ty, bao gồm Canary và Canarytokens, là những bẫy và dây cờ được thiết kế để phát hiện và ngăn chặn các kẻ tấn công mạng. Một tính năng quan trọng của Canarytokens là token khóa API AWS, rất hấp dẫn đối với các kẻ tấn công. Thinkst cũng đã hợp tác với một ngân hàng để cung cấp một token thẻ tín dụng thực. Công ty đã giới thiệu một tính năng mới gọi là Breadcrumbs, giúp hướng dẫn các kẻ tấn công đến các canary, cho phép phát hiện và phản hồi hiệu quả hơn. Thinkst đã đạt được sự tăng trưởng đáng kể, đạt 22,5 triệu đô la doanh thu định kỳ hàng năm (ARR) mà không cần bán hàng trực tiếp hoặc tăng giá trong thập kỷ qua. Tập cũng bao gồm một bản demo trực tiếp và thảo luận về các sản phẩm và dịch vụ của công ty.

TIPS & TRICKS CHO DEV

### Cài đặt Ollama

**Vấn đề:** Cần cài đặt Local LLM trên máy cá nhân.

**Cách làm:** Sử dụng lệnh `ollama install` để cài đặt, sau đó chạy `ollama init` để khởi tạo. Ví dụ: `ollama install ollama/llm-base`.

**Đánh giá:** Hiệu quả cao, nên dùng khi cần chạy LLM offline.

### Chạy LM Studio

**Vấn đề:** Cần chạy Local LLM trên máy cá nhân mà không cần internet.

**Cách làm:** Sử dụng lệnh `lm-studio run` để chạy, sau đó nhập prompt như "Hello, tôi muốn hỏi về thời tiết".

**Đánh giá:** Hiệu quả cao, nên dùng khi cần chạy LLM offline thường xuyên.

### Tối ưu hóa LLM

**Vấn đề:** Cần tối ưu hóa hiệu suất của Local LLM.

**Cách làm:** Sử dụng lệnh `ollama optimize` để tối ưu hóa, sau đó kiểm tra hiệu suất với lệnh `ollama benchmark`.

**Đánh giá:** Hiệu quả tốt, nên dùng khi cần cải thiện hiệu suất của LLM.

---

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để phát triển ứng dụng AI hiệu quả, dev cần biết cách tối ưu chi phí và hiệu năng của Large Language Model (LLM). Điều này giúp giảm thiểu chi phí vận hành và cải thiện thời gian phản hồi của ứng dụng. Việc tối ưu hóa cũng giúp tăng cường bảo mật và giảm thiểu rủi ro bảo mật.
3. Ví dụ, sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation of Large Language Models) có thể giúp giảm kích thước mô hình và tăng tốc độ xử lý. Ví dụ code: `model = LLaMA.from_pretrained("decapoda-research/llama-7b-hf"); model.fine_tune(...)`
4. Tip hoặc bước tiếp theo: Nghiên cứu và áp dụng các kỹ thuật tối ưu hóa LLM như pruning, quantization và knowledge distillation để giảm thiểu chi phí và cải thiện hiệu năng của ứng dụng AI.