BIS - MT · 06/08/2026

# Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

> *"Cứng quá thì gãy, mềm quá thì uốn."*
>
> — Tục ngữ Việt Nam

*Sự linh hoạt và cân bằng là chìa khóa — biết khi nào cứng rắn, khi nào nhường nhịn mới là trí khôn.*

## TIN TỨC NỔI BẬT

### 1 Hiện trạng 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 "Trạng thái của các Framework Trình điều khiển: Chọn Thời gian chạy Phù hợp cho Thực thi AI Doanh nghiệp" cung cấp cái nhìn tổng quan về trạng thái hiện tại của các framework trình điều khiển trong bối cảnh thực thi AI doanh nghiệp. Các framework trình điều khiển là các nền tảng phần mềm cho phép tích hợp trí tuệ nhân tạo (AI) và các model học máy (ML) vào các ứng dụng kinh doanh. Bài viết nhấn mạnh tầm quan trọng của việc chọn thời gian chạy phù hợp cho thực thi AI doanh nghiệp, vì nó có thể ảnh hưởng đáng kể đến hiệu suất, khả năng mở rộng và khả năng bảo trì của các ứng dụng được hỗ trợ bởi AI. Nó thảo luận về các framework trình điều khiển khác nhau, bao gồm Rasa, Dialogflow và Botpress, cũng như các điểm mạnh và điểm yếu tương ứng của chúng. Bài viết cũng đề cập đến các thách thức khi triển khai và bảo trì các ứng dụng được hỗ trợ bởi AI, chẳng hạn như tích hợp dữ liệu, cập nhật model, và các vấn đề về bảo mật. Nó nhấn mạnh nhu cầu về một framework trình điều khiển linh hoạt và có khả năng mở rộng có thể thích nghi với các yêu cầu kinh doanh thay đổi và tiến bộ công nghệ. Cuối cùng, bài viết nhằm cung cấp hướng dẫn cho các doanh nghiệp trong việc chọn framework trình điều khiển phù hợp nhất cho nhu cầu thực thi AI của họ, đảm bảo tích hợp liền mạch và hiệu suất tối ưu của các ứng dụng được hỗ trợ bởi AI.

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

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

marktechpost.com Đọc bài viết →

Amazon Web Services (AWS) đã thực hiện một bước tiến quan trọng trong lĩnh vực phát triển trí tuệ nhân tạo (AI) bằng cách mở mã nguồn một máy chủ MCP cho Bedrock AgentCore. Động thái này nhằm mục đích đơn giản hóa việc phát triển các tác nhân AI bằng cách cung cấp một giải pháp tiêu chuẩn hóa và mã nguồn mở. Máy chủ MCP được thiết kế để hoạt động với Bedrock AgentCore, một nền tảng cho phép phát triển các tác nhân AI. Bằng cách mở mã nguồn máy chủ MCP, AWS đang giúp các nhà phát triển xây dựng, thử nghiệm và triển khai các tác nhân AI dễ dàng hơn. Động thái này dự kiến sẽ đẩy nhanh việc phát triển các tác nhân AI và giảm độ phức tạp liên quan đến việc xây dựng và tích hợp các hệ thống AI. Việc mở mã nguồn máy chủ MCP là một bước tiến quan trọng trong lĩnh vực phát triển AI. Nó cho phép các nhà phát triển đóng góp vào cơ sở mã, sửa lỗi và thêm tính năng mới, khiến nó trở thành một giải pháp mạnh mẽ và đáng tin cậy hơn. Động thái này dự kiến sẽ có tác động tích cực đến cộng đồng phát triển AI, cho phép phát triển các tác nhân AI nhanh chóng và hiệu quả hơn.

## 3 Tăng tốc phát triển với MCP server của Amazon Bedrock AgentCore | AI

*Accelerate development with the Amazon Bedrock AgentCore MCP server | Artificial Intelligence*

Amazon Web Services (AWS) Đọc bài viết →

Amazon Web Services (AWS) đã giới thiệu máy chủ Amazon Bedrock AgentCore MCP, được thiết kế để tăng tốc phát triển trong lĩnh vực Trí tuệ Nhân tạo (AI). Máy chủ Bedrock AgentCore MCP là một thành phần chính của nền tảng Amazon Bedrock, nhằm mục đích đơn giản hóa việc phát triển và triển khai các model AI. Máy chủ này được tối ưu hóa cho tính toán hiệu suất cao và cung cấp một môi trường có thể mở rộng và bảo mật cho các nhà phát triển xây dựng và đào tạo các model AI. Với máy chủ Bedrock AgentCore MCP, các nhà phát triển có thể truy cập vào một loạt các công cụ và dịch vụ AI, bao gồm các framework học máy, lưu trữ dữ liệu và phân tích. Việc giới thiệu máy chủ Bedrock AgentCore MCP dự kiến sẽ cho phép phát triển AI nhanh

hơn và hiệu quả hơn, cho phép các nhà phát triển tập trung vào xây dựng các ứng dụng và giải pháp sáng tạo. Bằng cách cung cấp một cơ sở hạ tầng mạnh mẽ và có thể mở rộng, AWS đang trao quyền cho các nhà phát triển đẩy ranh giới của những gì có thể thực hiện được với AI và thúc đẩy đổi mới trong các ngành công nghiệp khác nhau.

4

## Điểm tin: Câu chuyện Startup Ấn Độ trong tuần

*News Roundup: Indian Startup Stories of the Week*

A Developer-Focused Guide  Đọc bài viết →

Cảnh quan khởi nghiệp của Ấn Độ đã tạo ra những điểm dữ liệu đáng kể, vòng vốn và sự thay đổi quy định. Một hướng dẫn gần đây đã chắt lọc những tiêu đề có thể hành động nhất từ "Tóm tắt Tin tức: Câu chuyện Khởi nghiệp Ấn Độ trong Tuần" của Dailyhunt. Hướng dẫn này nhằm chuyển đổi tin tức thành một tài sản quý giá cho các nhà phát triển, người sáng lập và những người xây dựng AI. Để đạt được điều này, một dịch vụ Python và FastAPI đã được tạo ra để kéo tin tức tài trợ mới nhất từ nguồn cấp RSS của Dailyhunt, lọc các khởi nghiệp Ấn Độ và gửi thông báo Slack. Dịch vụ này cung cấp một tài sản tổng hợp theo thời gian thực có thể được sử dụng trong các bảng điều khiển phân tích thị trường. Ngoài ra, Bộ Điện tử và Công nghệ Thông tin (MeitY) của Ấn Độ đã phát hành Dự luật Bảo vệ Dữ liệu Cá nhân (PDP) năm 2024, nhấn mạnh tầm quan trọng của việc tuân thủ các quy định bảo vệ dữ liệu. Nhúng các mẫu này có thể giúp một ngăn xếp được bảo vệ trong tương lai khỏi các hình phạt về tuân thủ và cung cấp một tín hiệu tin cậy cho các nhà đầu tư.

5

## Meta tham gia cuộc chiến AI coding với Muse Spark 1.2 và Muse Code cùng các async background agent hoạt động liên tục

*Meta enters the AI coding wars with Muse Spark 1.2 and Muse Code with persistent async background agents*

VentureBeat  Đọc bài viết →

Meta đã tham gia vào cuộc chiến mã hóa AI với việc phát hành Muse Code, một đại lý mã hóa AI dựa trên□□ trong phiên bản beta, và Muse Spark 1.2, một bản cập nhật tập trung vào mã hóa cho dòng model Muse Spark tiên phong. Muse Code được thiết kế để thực hiện các nhiệm vụ kỹ thuật phần mềm hoàn chỉnh, bao gồm lập kế hoạch, viết mã, và xác thực kết quả, và có thể được cài đặt trên macOS hoặc Linux với một lệnh curl duy nhất. Đại lý này sử dụng các đại lý nền

tảng bền bỉ, vẫn hoạt động trong suốt mỗi phiên, để giảm độ trễ và cải thiện hiệu suất. Muse Spark 1.2 là một model độc quyền đã được đào tạo cùng với Muse Code và được thiết kế để cải thiện việc tạo mã, gỡ lỗi phức tạp, và hiểu biết về cơ sở mã. Model này đã được thử nghiệm trên các điểm chuẩn khác nhau và đã thể hiện sự cải thiện đáng kể so với người tiền nhiệm của nó, Muse Spark 1.1. Meta đang cung cấp Muse Spark 1.2 thông qua API Model Meta của mình trong hai cấp độ: tiêu chuẩn và người đóng góp. Cấp độ tiêu chuẩn được định giá 1,25 đô la cho mỗi triệu token đầu vào và 4,25 đô la cho mỗi triệu token đầu ra, trong khi cấp độ người đóng góp được định giá 0,10 đô la cho mỗi triệu token đầu vào và 0,20 đô la cho mỗi triệu token đầu ra, nhưng yêu cầu các nhà phát triển phải cung cấp thông tin thanh toán và đăng nhập bằng tài khoản Meta. Cấp độ người đóng góp đã gây ra lo ngại trong số các nhà phát triển và doanh nghiệp muốn giữ mã của họ an toàn, vì nó yêu cầu họ phải tài trợ cho việc truy cập bằng dữ liệu của riêng họ. Quyết định của Meta chuyển khỏi các model mã nguồn mở, vốn là một phần quan trọng trong chiến lược của họ trong quá khứ, cũng đã bị chỉ trích. Việc phát hành Muse Code và Muse Spark 1.2 đánh dấu sự tham gia nghiêm túc nhất của Meta vào cuộc chiến mã hóa AI cho đến nay, và sẽ rất thú vị khi xem nó cạnh tranh với Claude Code của Anthropic và Codex của OpenAI.

6

## Đánh giá an ninh mạng từ bên thứ ba liên quan đến các model của OpenAI

*Third-party cyber evaluations involving OpenAI models*

OpenAI Blog  Đọc bài viết →

OpenAI đã giải quyết các vụ việc gần đây liên quan đến việc đánh giá an ninh mạng của bên thứ ba đối với các model AI của họ. Công ty đã thực hiện các biện pháp để tăng cường bảo mật và tính toàn vẹn của các model của mình thông qua việc triển khai các biện pháp phòng vệ mới.

7

## Giới thiệu Muse Code và Muse Spark 1.2

*Introducing Muse Code and Muse Spark 1.2*

Simon Willison  Đọc bài viết →

Meta đã phát hành bản cập nhật cho mô hình Muse Spark của mình, phiên bản 1.2, tập trung vào khả năng lập trình. Bản cập nhật này, kết

hợp với việc giới thiệu Muse Code, nhằm cải thiện khả năng gọi công cụ đại lý trình tự dài. Muse Spark 1.2 có các tính năng cải tiến trong việc tạo mã, gỡ lỗi, hiểu cơ sở mã và luồng công việc của nhà phát triển từ đầu đến cuối. Mô hình này đã được đào tạo trên quy mô lớn hơn với sự đa dạng tăng lên trong môi trường đào tạo, và hiệu suất của nó đã được tối ưu hóa để sử dụng với Muse Code. Bản cập nhật bao gồm đào tạo rộng rãi về các nhiệm vụ lập trình đường chân trời dài, chẳng hạn như tạo toàn bộ kho lưu trữ và các dự án từ đầu đến cuối lớn. Phát hành này thể hiện sự đầu tư liên tục của Meta vào việc phát triển các mô hình AI cho lập trình và gọi công cụ.

8

## Đội ngũ pháp lý của GitHub đã sử dụng Copilot CLI để tinh gọn workflow của họ như thế nào

*How the GitHub legal team used Copilot CLI to streamline their workflows*

GitHub Blog  Đọc bài viết →

Đội ngũ pháp lý của GitHub đã thành công trong việc tận dụng công cụ Copilot CLI để tự động hóa quy trình làm việc và tăng năng suất. Đội ngũ này, bao gồm các luật sư, quản lý chương trình và chuyên gia kinh doanh, phải đối mặt với các nhiệm vụ lặp đi lặp lại như xem xét hợp đồng và trả lời các câu hỏi pháp lý. Bằng cách tận dụng Copilot CLI, họ đã có thể tự động hóa các nhiệm vụ này và xây dựng các công cụ tùy chỉnh mà không cần kiến thức lập trình rộng rãi. Đội ngũ chỉ cần mô tả kết quả mong muốn của họ bằng ngôn ngữ đơn giản, tích hợp Copilot CLI vào kho lưu trữ của họ và chứng kiến kết quả ngay lập tức. Sự thay đổi này trong quy trình làm việc đã cho phép các thành viên không kỹ thuật xây dựng các giải pháp tùy chỉnh, thúc đẩy văn hóa đổi mới và hợp tác. Kết quả là, năng suất và hiệu quả của đội ngũ đã được cải thiện, cho phép họ tập trung vào các nhiệm vụ phức tạp hơn.

TIPS & TRICKS CHO DEV

### Cài đặt AI Observability

**Vấn đề:** Thiếu thông tin về hiệu suất của mô hình AI.

**Cách làm:** Sử dụng LangSmith để theo dõi hiệu suất của mô hình AI. Ví dụ, `langsmith monitor --model my_model` để bắt đầu theo dõi.

**Đánh giá:** Hiệu quả trong việc tối ưu hóa mô hình AI, nên dùng khi triển khai mô hình vào sản xuất.

### Thực hiện Tracing

**Vấn đề:** Khó tìm ra nguyên nhân của lỗi trong mô hình AI.

**Cách làm:** Sử dụng Langfuse để thực hiện tracing, ví dụ `langfuse trace --model my_model --input my_input`.

**Đánh giá:** Giúp tìm ra nguyên nhân lỗi nhanh chóng, nên dùng khi gặp lỗi không rõ nguyên nhân.

### Kiểm soát Chi Phí

**Vấn đề:** Chi phí đào tạo và triển khai mô hình AI quá cao.

**Cách làm:** Sử dụng Arize Phoenix để theo dõi và kiểm soát chi phí, ví dụ `arize phoenix --model my_model --cost-budget 1000`.

**Đánh giá:** Giúp giảm chi phí và tối ưu hóa tài nguyên, nên dùng khi có ngân sách hạn chế.

---

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tối ưu chi phí & hiệu năng LLM**
2. Để xây dựng ứng dụng hiệu quả, các nhà phát triển cần biết cách tối ưu hóa chi phí và hiệu năng của mô hình ngôn ngữ lớn (LLM). Điều này giúp giảm thiểu chi phí tính toán và tăng tốc độ xử lý, từ đó cải thiện trải nghiệm người dùng. Việc tối ưu hóa LLM cũng giúp giảm thiểu tác động môi trường của ứng dụng.
3. Ví dụ, việc sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) có thể giúp giảm kích thước mô hình và tăng tốc độ đào tạo, từ đó giảm chi phí và tăng hiệu năng.
4. Để tiếp tục, bạn có thể nghiên cứu về các kỹ thuật tối ưu hóa LLM khác như Quantization, Pruning và Knowledge Distillation để áp dụng vào dự án của mình.