BIS - MT · 18/07/2026

# Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

> *“Sóng cả không lo, sóng lớn không sợ.”*
>
> — Tục ngữ Việt Nam

*Người có bản lĩnh không bị dao động trước khó khăn — sự bình thản và dũng cảm là sức mạnh thật sự.*

## TIN TỨC NỔI BẬT

### 1 Xây dựng agent phân tích tài chính thông minh với LangGraph và Strands Agents

*Build an intelligent financial analysis agent with LangGraph and Strands Agents*

Amazon Web Services (AWS) Đọc bài viết →

Amazon Web Services (AWS) đã giới thiệu một khả năng mới để xây dựng một tác nhân phân tích tài chính thông minh sử dụng LangGraph và Strands Agents. Giải pháp sáng tạo này cho phép các doanh nghiệp tạo ra các công cụ phân tích tài chính tinh vi có thể xử lý và phân tích lượng lớn dữ liệu tài chính. LangGraph, một công nghệ xử lý ngôn ngữ tự nhiên (NLP), được sử dụng để trích xuất thông tin tài chính liên quan từ các nguồn dữ liệu không cấu trúc như báo cáo tài chính, bài viết tin tức và bài đăng trên mạng xã hội. Strands Agents, một nền tảng low-code, được sử dụng để xây dựng và triển khai tác nhân phân tích tài chính. Với giải pháp này, các doanh nghiệp có thể tự động hóa các nhiệm vụ phân tích tài chính, chẳng hạn như xác định xu hướng, phát hiện bất thường và dự đoán kết quả tài chính. Tác nhân thông minh cũng có thể cung cấp thông tin và khuyến nghị tài chính được cá nhân hóa cho người dùng. Bằng cách tận dụng LangGraph và Strands Agents, các doanh nghiệp có thể có được hiểu biết sâu sắc hơn về hiệu suất tài chính của mình và đưa ra quyết định thông minh hơn. Giải pháp này đặc biệt hữu ích cho các tổ chức tài chính, công ty kế toán và các doanh nghiệp yêu cầu khả năng phân tích tài chính tiên tiến.

### 2 Các model mã nguồn mở 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 →

> Thật không may, nội dung được cung cấp chỉ là tiêu đề và tên trang web, không có nội dung thực sự. Tôi không thể viết tóm tắt dựa trên thông tin đã cho. Nếu bạn có thể cung cấp nội dung thực sự, tôi sẽ rất vui được giúp bạn viết một bản tóm tắt rõ ràng và đầy đủ thông tin.

## 3 MiniMax-M2 là "vua" mới của các LLM mã nguồn mở (đặc biệt cho agentic tool calling)

*MiniMax-M2 is the new king of open source LLMs (especially for agentic tool calling)*

Venturebeat Đọc bài viết →

> MiniMax-M2 đã nổi lên như mô hình ngôn ngữ lớn (LLM) mã nguồn mở hàng đầu, đặc biệt là vượt trội trong việc gọi công cụ đại lý. Sự tiến bộ này trong lĩnh vực LLM đã thu hút sự chú ý của cộng đồng công nghệ. Khả năng của mô hình trong việc gọi công cụ đại lý, bao gồm tương tác với các công cụ và hệ thống để thực hiện nhiệm vụ, đã giúp nó khác biệt so với các LLM mã nguồn mở khác. Hiệu suất của MiniMax-M2 trong việc gọi công cụ đại lý rất đáng chú ý, thể hiện tiềm năng của nó trong việc tự động hóa các nhiệm vụ phức tạp và nâng cao năng suất. Là một mô hình mã nguồn mở, nó có thể truy cập được đối với các nhà phát triển và nhà nghiên cứu, cho phép họ xây dựng và cải thiện khả năng của nó. Sự cởi mở này thúc đẩy sự hợp tác và đổi mới, thúc đẩy tiến bộ trong lĩnh vực LLM. Sự xuất hiện của MiniMax-M2 như một mô hình LLM mã nguồn mở mới nổi bật cho thấy sự tiến bộ nhanh chóng trong lĩnh vực này. Hiệu suất vượt trội của nó trong việc gọi công cụ đại lý đã đặt nó vào vị trí một công cụ quý giá cho các nhà phát triển và tổ chức muốn tận dụng sức mạnh của LLM trong các ứng dụng của họ.

## 4 Đo lường ROI của AI trong tài chính: LangSmith và Pay-i giúp kết nối chi phí và giá trị

LangChain Blog Đọc bài viết →

> Các tổ chức dịch vụ tài chính đang gặp khó khăn trong việc đo lường lợi tức đầu tư (ROI) của chi tiêu cho trí tuệ nhân tạo (AI). Mặc dù các hệ thống đa tác nhân đã thành công trong việc tự động hóa các nhiệm vụ như xử lý RFP và giám sát tuân thủ chống rửa tiền (AML), nhưng

nhiều đội vẫn không thể lượng hóa được lợi ích của các hệ thống này. Vấn đề nằm ở kinh tế phức tạp của các hệ thống này, liên quan đến các phương trình động, đa biến mà các công cụ hoạt động tài chính truyền thống không được trang bị để xử lý. Để giải quyết vấn đề này, hai nền tảng, LangSmith và Pay-i, đã được phát triển để cung cấp cho các đội kỹ sư cái nhìn sâu vào cấu trúc chi phí của các hệ thống đa tác nhân và đo lường giá trị kinh doanh của chúng. LangSmith theo dõi chi phí của các cuộc gọi tác nhân, bao gồm cả cuộc gọi LLM, cuộc gọi công cụ và chi phí điều phối, trong khi Pay-i kết nối chi phí của các trường hợp sử dụng GenAI với các kết quả kinh doanh có thể đo lường được, chẳng hạn như thời gian tiết kiệm hoặc giá trị tạo ra. Bằng cách sử dụng cả hai nền tảng này cùng nhau, các tổ chức dịch vụ tài chính có thể đóng vòng lặp giữa chi phí và giá trị của các hệ thống AI của họ, cung cấp một câu trả lời rõ ràng cho câu hỏi về việc họ nhận lại được gì từ khoản đầu tư của mình.

5

## Ba điểm yếu trong bảo mật doanh nghiệp ở quy mô codebase (và lý do các công cụ hiện tại không giải quyết được)

*Three places enterprise security breaks down at codebase scale (and why your current tools don't cover them)*

Sourcegraph Blog  Đọc bài viết →

Bảo mật doanh nghiệp thường gặp thách thức ở quy mô cơ sở mã, nơi các điểm yếu trong mã phần mềm gây ra rủi ro đáng kể. Một lĩnh vực quan trọng mà bảo mật bị phá vỡ là trong việc khắc phục điểm yếu. Quá trình này liên quan đến việc xác định và sửa lỗi điểm yếu mã, nhưng nó trở nên ngày càng khó khăn khi đối phó với các cơ sở mã lớn, chẳng hạn như những cơ sở mã được tìm thấy trong 10.000 kho lưu trữ. Vấn đề này còn được phức tạp hóa bởi việc sử dụng mã được tạo bởi AI, điều này có thể khó phân tích và hiểu nếu không có ngữ cảnh. Ngoài ra, khoảng trống trong phạm vi phát hiện có nghĩa là một số điểm yếu có thể không được phát hiện bởi các công cụ bảo mật hiện tại, khiến chúng không được giải quyết. Quá trình khắc phục bản thân cũng là một thách thức, vì việc sửa lỗi có thể mất vài tuần để thực hiện, cho phép các điểm yếu vẫn mở trong thời gian dài. Điều này nhấn mạnh nhu cầu về các giải pháp bảo mật hiệu quả và hiệu suất hơn có thể xử lý các phức tạp của các cơ sở mã lớn và mã được tạo bởi AI.

## Sau Orthogonality: Agency đạo đức đức hạnh và AI Alignment

6

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

The Gradient  Đọc bài viết →

Bài viết "Sau Tính Chéo: 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 của các hành động, khuynh hướng hành động, tiêu chí đánh giá và tài nguyên mà cấu trúc và thúc đẩy bản thân. Tác giả đề xuất rằng AI nên sử dụng một "chữ ký loại" tương tự như logic dựa trên thực hành của con người để hỗ trợ, cộng tác hoặc tuân thủ 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 khám phá khái niệm eudaimonia, hay sự thịnh vượng tích cực, hợp lý của con người, và cho rằng nó chỉ ra một cấu trúc suy nghĩ khác với tính hợp lý kết quả tiêu chuẩn. Tác giả đề xuất tính hợp lý eudaimonic như một khuôn khổ hữu ích cho AI liên kết với con người, bao gồm việc đánh giá cao các hành động như các yếu tố của một thực hành chứ không phải như một phương tiện để đạt được mục tiêu. Cách tiếp cận này được coi là ổn định và an toàn hơn so với cơ quan deontological và kết quả. Tác giả cho rằng tính hợp lý eudaimonic là một hình thức cơ quan tự nhiên có thể hiệu quả trong việc liên kết AI với sự thịnh vượng của con người. Họ lập luận rằng cách tiếp cận này có thể giải quyết các vấn đề an toàn AI cổ điển và nghịch lý, khiến nó trở thành một hướng đi đầy hứa hẹn cho sự liên kết AI.

7

## Giờ đây, ngay cả các hacker tinh nhuệ nhất của Nga cũng dùng Clickfix để lây nhiễm thiết bị

*Now, even Russia's most elite hackers are using Clickfix to infect devices*

Ars Technica  Đọc bài viết →

Nhóm hack tinh nhuệ của Nga, Sandworm, đã áp dụng một kỹ thuật kỹ thuật xã hội gọi là Clickfix để xâm phạm các thiết bị thuộc các tổ chức nhạy cảm ở Ukraine. Clickfix liên quan đến việc hiển thị một CAPTCHA trên một trang web yêu cầu khách truy cập sao chép và dán một hỗn hợp văn bản vào terminal, chứa các tập lệnh độc hại có thể cài đặt malware hoặc trích xuất dữ liệu nhạy cảm. Kỹ thuật này đã được sử dụng bởi các tội phạm có động cơ tài chính trong quá khứ, nhưng Sandworm, một đơn vị hack tiên tiến trong cánh tay tình báo quân sự của Nga, đã áp dụng nó. Các cơ quan chức năng Ukraine đã phát hiện 10 trang web bị xâm phạm hiển thị một lệnh PowerShell như một phần của CAPTCHA giả, có thể cài đặt các tập lệnh Visual Basic độc hại và malware khác. Malware được sử dụng trong chiến dịch bao

gồm FreakyPoll, một tập lệnh Python tạo cửa hậu cho thiết bị, và FluidLeech, một chương trình ngụy trang thành phần mềm chống vi-rút. Các cuộc tấn công đã dẫn đến việc xâm phạm mạng của ít nhất một tổ chức, với malware thu thập thông tin và tạo cửa hậu cho hệ thống.

8

## Cách các harness tự cải thiện đang viết lại playbook kỹ thuật agent

*How self-improving harnesses are rewriting the agent engineering playbook*

BD Tech Talks Đọc bài viết →

Các ứng dụng trí tuệ nhân tạo (AI) phụ thuộc nặng nề vào khung chạy thời gian thực của chúng, bao gồm logic thực thi, lời nhắc hệ thống, quản lý bộ nhớ và cấu hình công cụ. Tuy nhiên, việc cập nhật và tối ưu hóa thủ công các khung này cho mỗi mô hình AI mới là một quá trình tốn thời gian và lao động. Các khuôn khổ AI gần đây đang giải quyết thách thức này bằng cách cho phép các tác nhân AI phân tích, thử nghiệm và tối ưu hóa môi trường thời gian chạy của chính chúng. Điều này được thực hiện thông qua một khung tự cải tiến cấu trúc hệ điều hành của tác nhân, cho phép nó tinh chỉnh hiệu suất một cách lặp đi lặp lại. Một ví dụ chính của công nghệ này là Claude Code của Anthropic, gần đây đã bị rò rỉ và phân tích bởi các nhà nghiên cứu bảo mật. Kiến trúc của Claude Code tách kế hoạch khỏi thực thi, sử dụng một tác nhân chính để phân tích yêu cầu và các tác nhân phụ chuyên dụng để xử lý thử nghiệm, tài liệu và gỡ lỗi song song. Tuy nhiên, khi các nhà phát triển muốn tối ưu hóa một tác nhân cho một ứng dụng cụ thể, họ phải điều chỉnh khung tại các mức độ khác nhau, điều này có thể là một quá trình thủ công và dễ xảy ra lỗi. Một khuôn khổ mới gọi là Self-Harness đã được phát triển để giải quyết thách thức này. Nó giới thiệu một vòng lặp tự động, lặp đi lặp lại cho phép các tác nhân AI cải thiện chính khung của chúng bằng cách khai thác dấu vết thực thi. Khuôn khổ này hoạt động ở ba giai đoạn: khai thác điểm yếu, đề xuất khung và xác thực đề xuất. Khi được thử nghiệm trên Terminal-Bench-2.0, vòng lặp Self-Harness đã tạo ra các quy tắc thực thi mới giúp cải thiện đáng kể hiệu suất của mô hình AI, đạt tỷ lệ vượt qua 61,9% trên các tiêu chuẩn chuẩn.

TIPS & TRICKS CHO DEV

### Quản lý context window

**Vấn đề:** Các mô hình AI thường bị giới hạn bởi kích thước context window.
**Cách làm:** Sử dụng kỹ thuật chunking để chia nhỏ văn bản, ví dụ prompt "Summarize the following text in 50 words: ...".
**Đánh giá:** Hiệu quả khi xử lý văn bản ngắn, nhưng có thể mất thông tin đối với văn bản dài.

### Tối ưu hóa long-context

**Vấn đề:** Các mô hình AI thường gặp khó khăn khi xử lý văn bản dài.
**Cách làm:** Sử dụng kỹ thuật attention mechanism để tập trung vào các phần quan trọng, ví dụ lệnh CLI `--max_length 2048`.
**Đánh giá:** Hiệu quả khi xử lý văn bản dài, nhưng có thể tiêu tốn nhiều tài nguyên.

### Sử dụng memory hiệu quả

**Vấn đề:** Các mô hình AI thường tiêu tốn nhiều bộ nhớ khi xử lý dữ liệu lớn.
**Cách làm:** Sử dụng kỹ thuật caching để lưu trữ dữ liệu tạm thời, ví dụ prompt "Use cached knowledge to answer: ...".
**Đánh giá:** Hiệu quả khi xử lý dữ liệu lớn, nhưng có thể mất cập nhật nếu dữ liệu thay đổi.

---

BÀI HỌC AI HÔM NAY CHO DEV

1. **Tích hợp AI API vào ứng dụng**
Dev cần biết về tích hợp AI API để tăng cường khả năng thông minh cho ứng dụng của mình, giúp tự động hóa các tác vụ và cải thiện trải nghiệm người dùng. Điều này đòi hỏi kiến thức về các thư viện và framework AI, cũng như cách tích hợp chúng vào dự án.

Ví dụ, sử dụng thư viện như TensorFlow hoặc PyTorch để tích hợp mô hình học máy vào ứng dụng, cho phép thực hiện các nhiệm vụ như nhận diện hình ảnh hoặc phân tích văn bản.

Tip hoặc bước tiếp theo: Hãy bắt đầu bằng việc nghiên cứu các thư viện AI phổ biến và chọn một phù hợp với nhu cầu dự án, sau đó thực hiện tích hợp và tinh chỉnh mô hình để đạt hiệu suất tốt nhất.