

Bản Tin AI Hằng Ngày

Cập nhật công nghệ AI mới nhất

“Một câu nhin là chín câu lành.”

— Tục ngữ Việt Nam

Biết kiềm chế và nhường nhịn trong lúc bức bối sẽ giữ được hòa khí và tránh hối hận về sau.

TIN TỨC NỔI BẬT

1

Giới thiệu Open Agent Specification (Agent Spec): Đại diện 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. Tiêu chuẩn này nhằm cung cấp một khuôn khổ chung cho các nhà phát triển để tạo và tích hợp các tác nhân AI trên nhiều nền tảng và ứng dụng khác nhau. Agent Spec được thiết kế để mở và có thể mở rộng, cho phép các nhà phát triển dễ dàng thêm các tính năng và khả năng mới. Tiêu chuẩn này định nghĩa một tập hợp các giao diện và cấu trúc dữ liệu tiêu chuẩn cho các tác nhân AI, cho phép giao tiếp và tương tác liền mạch giữa các tác nhân và hệ thống khác nhau. Biểu diễn thống nhất này dự kiến sẽ thúc đẩy sự phát triển của các ứng dụng AI phức tạp và tích hợp hơn. Bằng cách cung cấp một ngôn ngữ và khuôn khổ chung cho các tác nhân AI, Agent Spec có tiềm năng đẩy nhanh sự đổi mới và áp dụng các công nghệ AI trên nhiều ngành công nghiệp. Các nhà phát triển có thể tận dụng Agent Spec để tạo ra các giải pháp AI hiệu quả, có thể mở rộng và tương tác hơn, cuối cùng thúc đẩy giá trị kinh doanh và cải thiện trải nghiệm khách hàng. Open Agent Specification hiện có sẵn cho các nhà phát triển để khám phá và sử dụng trong các dự án AI của họ.

2

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 đã giới thiệu Agent Factory, một nền tảng kết nối các tác nhân, ứng dụng và dữ liệu bằng cách sử dụng các tiêu chuẩn mở mới. Nền tảng này tận dụng Microsoft Cloud PC (MCP) và Azure Application API (A2A) để tạo điều kiện cho sự tương tác liền mạch giữa các hệ thống khác nhau. Agent Factory cho phép các nhà phát triển tạo và tích hợp các tác nhân tùy chỉnh với các ứng dụng và nguồn dữ liệu khác nhau, nâng cao trải nghiệm người dùng tổng thể. Các tiêu chuẩn mở của nền tảng cho phép linh hoạt và khả năng tương tác cao hơn, giúp các nhà phát triển xây dựng và triển khai các ứng dụng đáp ứng nhu cầu kinh doanh cụ thể. Bằng cách tận dụng MCP và A2A, Agent Factory cung cấp một môi trường có thể mở rộng và bảo mật cho các tác nhân hoạt động. Điều này cho phép các tổ chức tự động hóa các nhiệm vụ, cải thiện hiệu quả và đưa ra quyết định dựa trên dữ liệu. Các tiêu chuẩn mở của nền tảng cũng tạo điều kiện cho sự hợp tác và đổi mới giữa các nhà phát triển, cho phép họ xây dựng trên các giải pháp hiện có và tạo ra các giải pháp mới. Agent Factory là một phần của Microsoft Azure, cung cấp một bộ công cụ và dịch vụ toàn diện để xây dựng và triển khai các ứng dụng dựa trên đám mây.

3

Tăng tốc phát triển với Amazon Bedrock AgentCore MCP server | Artificial Intelligence

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 này là một thành phần chính của nền tảng Amazon Bedrock, cung cấp dịch vụ quản lý để xây dựng, triển khai và quản lý các model AI. Máy chủ Amazon Bedrock AgentCore MCP là một thể hiện tính toán hiệu suất cao (HPC) cho phép các nhà phát triển chạy các khối lượng công việc AI quy mô lớn, chẳng hạn như đào tạo model và suy luận. Với máy chủ này, các nhà phát triển có thể xử lý nhanh chóng và hiệu quả các nhiệm vụ AI phức tạp, giảm thời gian và tài nguyên cần thiết cho phát triển. Máy chủ AgentCore MCP được tối ưu hóa cho các khối lượng công việc AI và cung cấp một môi trường có thể mở rộng và bảo mật cho các nhà phát triển xây dựng và triển khai các model AI. Bằng cách tận dụng máy chủ này, các nhà phát triển có thể tăng tốc quá trình phát triển AI của mình, cải thiện độ chính xác của model và giảm chi phí liên quan đến phát triển và triển khai AI.

Máy chủ Amazon Bedrock AgentCore MCP là một công cụ quý giá cho các tổ chức muốn tận dụng sức mạnh của AI và học máy trong các ứng dụng của họ.

4

Khi rủi ro sự nghiệp đáng để chấp nhận

When Career Risks Are Worth Taking

IEEE Spectrum

[Đọc bài viết →](#)

Các kỹ sư phần mềm có thời gian gắn bó ngắn ngủi, với thời gian trung bình khoảng hai năm tại một công ty, thấp hơn đáng kể so với các nghề nghiệp tri thức khác. Sự bất ổn này đã được nhấn mạnh bởi các đợt sa thải gần đây, nhưng nó cũng có thể là một cơ hội quý giá cho sự phát triển sự nghiệp. Tác giả, Brian Jenney, người đã chuyển sang nghề kỹ sư phần mềm vào tuổi 30, đã học cách quản lý rủi ro sự nghiệp một cách có chủ đích. Ông xác định hai loại rủi ro: rủi ro liên quan đến công việc và rủi ro liên quan đến xu hướng. Rủi ro liên quan đến công việc bao gồm việc tìm kiếm các vai trò và cơ hội tốt hơn, điều này có thể dẫn đến sự tăng cường chuyên môn và khả năng cạnh tranh. Tuy nhiên, cách tiếp cận này có thể trở nên căng thẳng và kém hiệu quả theo thời gian. Loại rủi ro thứ hai liên quan đến việc làm việc cùng với các kỹ sư mạnh hơn, điều này có thể thách thức nhưng dẫn đến việc hấp thụ các kỹ năng và chuyên môn quý giá. Cách tiếp cận này có thể tạo ra sự ổn định lâu dài và tăng khả năng cạnh tranh. Bài viết cũng đề cập đến tầm quan trọng của việc đặt cược vào các xu hướng như dịch vụ đám mây, ReactJS và AI để duy trì sự liên quan trong ngành. Bằng cách chấp nhận rủi ro có tính toán, các kỹ sư phần mềm có thể biến sự bất ổn thành lợi thế cho sự nghiệp, chẳng hạn như phát triển các mô hình LLM mới, tạo ra các API mạnh mẽ, hoặc xây dựng các framework hiệu quả cho các nhà phát triển.

5

AGI không phải là Multimodal

AGI Is Not Multimodal

The Gradient

[Đọc bài viết →](#)

Những tiến bộ gần đây trong các mô hình AI tạo sinh đã khiến một số người tin rằng Trí tuệ Tổng quát Nhân tạo (AGI) đang sắp xảy ra. Tuy nhiên, những mô hình này đã xuất hiện không phải do các giải pháp sâu sắc cho vấn đề trí tuệ, mà vì chúng đã được mở rộng hiệu quả trên phần cứng hiện có. Phương pháp đa mô thức, kết hợp nhiều mô thức

để tạo ra một AI tổng quát, được coi là con đường dẫn đến AGI. Tuy nhiên, chiến lược này không thể thành công trong thời gian gần và có thể không dẫn đến AGI cấp độ con người có thể thực hiện các nhiệm vụ như lý luận cảm biến và phối hợp xã hội. Một cách tiếp cận hiệu quả hơn đến AGI có thể liên quan đến việc coi sự thể hiện và tương tác với môi trường là chính, và xem quá trình xử lý tập trung vào mô thức như hiện tượng xuất hiện. Điều này đòi hỏi một AGI thực sự phải tổng quát trên tất cả các lĩnh vực, bao gồm cả thực tại vật lý, nơi nó có thể giải quyết các vấn đề như sửa chữa ô tô hoặc chuẩn bị thức ăn. Sự hiểu biết về thế giới vật lý là cần thiết cho những vấn đề này, và các Mô hình Ngôn ngữ Lớn (LLM) hiện tại có thể không sở hữu sự hiểu biết này, mặc dù khả năng dự đoán token và phản ánh sự hiểu biết bề mặt về thực tại.

6

Đánh giá OneOdio Studio Max 2: Tai nghe độ trễ cực thấp dành cho tech enthusiast

OneOdio Studio Max 2 review: Ultra-low latency headphones for tech enthusiasts

BD Tech Talks

[Đọc bài viết →](#)

Tai nghe OneOdio Studio Max 2 là một đôi tai nghe không dây, che tai, thiết kế cho những người đam mê công nghệ, nhà phát triển và game thủ yêu cầu độ trễ thấp và kết nối linh hoạt. Những chiếc tai nghe này kết hợp nhiều chế độ không dây và có dây, bao gồm một bộ phát 2,4GHz chuyên dụng, để giảm thiểu độ trễ âm thanh. Chúng có thân nhựa đen mờ với headband điều chỉnh và earpad bọt biển nhớ để mang lại sự thoải mái. Tai nghe hỗ trợ bốn phương pháp kết nối, bao gồm Bluetooth 6.0, bộ phát USB M2 và âm thanh có dây qua jack cắm 3,5mm và 6,35mm. Chúng cũng hỗ trợ phát trực tuyến không dây độ phân giải cao với codec Sony LDAC. Tai nghe có trình điều khiển High-Flux 45mm và một ứng dụng đồng hành để thiết lập EQ và giới hạn âm lượng. Tuy nhiên, chúng thiếu Chế độ hủy tiếng ồn chủ động và chế độ minh bạch, dựa vào cách ly thụ động để giảm tiếng ồn. Studio Max 2 cung cấp hiệu suất pin vượt trội, với thời gian phát lại liên tục lên đến 120 giờ qua Bluetooth, và có giá bán lẻ 199,99 đô la.

7

Ba điểm yếu trong bảo mật enterprise ở quy mô codebase (và lý do các tool 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 bị thỏa hiệp ở quy mô codebase do một số vấn đề chính. Một lĩnh vực chính nơi bảo mật bị phá vỡ là trong việc khắc phục lỗ hổng. Quá trình này liên quan đến việc xác định và sửa lỗi lỗ hổng trong code, nhưng nó trở nên ngày càng thách thức khi xử lý các codebase lớn trải rộng hàng nghìn kho lưu trữ. Vấn đề này trở nên trầm trọng hơn do sử dụng code AI, điều này có thể thiếu ngữ cảnh và khiến các công cụ bảo mật khó phát hiện chính xác lỗ hổng. Ngoài ra, khoảng trống trong phạm vi phát hiện có thể xảy ra, khiến một số lỗ hổng không được phát hiện và không được giải quyết. Hơn nữa, thời gian cần thiết để sửa lỗi những lỗ hổng này có thể đáng kể, với các bản sửa lỗi thường mất vài tuần để triển khai trên một codebase lớn. Những vấn đề này làm nổi bật những hạn chế của các công cụ bảo mật hiện tại, có thể không được trang bị để xử lý quy mô và độ phức tạp của các codebase hiện đại. Do đó, bảo mật doanh nghiệp thường bị thỏa hiệp, khiến các tổ chức dễ bị tổn thương trước các mối đe dọa tiềm năng.

8

Chi phí của việc nói "có" đã thay đổi

The cost of saying yes has changed

GitHub Blog [Đọc bài viết →](#)

Chi phí phát triển phần mềm đã trải qua một sự thay đổi đáng kể với sự xuất hiện của trí tuệ nhân tạo (AI) và học máy. Lịch sử cho thấy, phần tốn kém nhất của một yêu cầu tính năng nhỏ là việc viết mã code. Tuy nhiên, với việc tạo mã code bằng AI, điều này đã thay đổi. Chi phí viết mã code đã giảm, nhưng chi phí sở hữu nó vẫn không thay đổi. Sự thay đổi này phá vỡ trực giác kỹ thuật truyền thống, nơi các kỹ sư sẽ xem xét kỹ lưỡng tính khả thi và tác động của một thay đổi trước khi thực hiện nó. Các kỹ sư hiện đang đối mặt với một thách thức mới: quyết định những thay đổi nào thực sự rẻ trong kỷ AI. Phần tốn kém nhất của một yêu cầu tính năng nhỏ không còn là viết mã code, mà là các cuộc họp và thảo luận để xác định có nên thực hiện nó hay không. Điều này đòi hỏi một khuôn khổ mới để đánh giá các thay đổi và quyết định những thay đổi nào đáng được theo đuổi. Bằng cách hiểu sự thay đổi này, các nhà phát triển có thể thích nghi phương pháp tiếp cận của mình với phát triển phần mềm và đưa ra quyết định thông minh hơn về những thay đổi nào nên ưu tiên.

TIPS & TRICKS CHO DEV

Tối ưu hóa mã

Vấn đề: Mã nguồn không tối ưu dẫn đến hiệu suất kém.

Cách làm: Sử dụng Claude Code với lệnh `claude optimize` để cải thiện mã. Ví dụ: `claude optimize --language python --input mycode.py`.

Đánh giá: Hiệu quả khi cải thiện mã, nhưng nên xem xét từng trường hợp cụ thể.

Tìm lỗi mã

Vấn đề: Khó tìm lỗi trong mã nguồn phức tạp.

Cách làm: Sử dụng Claude Code với lệnh `claude debug` để tìm lỗi. Ví dụ: `claude debug --language java --input mycode.java`.

Đánh giá: Hiệu quả khi tìm lỗi nhanh, nhưng nên kiểm tra kỹ lưỡng.

Refactor mã

Vấn đề: Mã nguồn không sạch sẽ và khó bảo trì.

Cách làm: Sử dụng Claude Code với lệnh `claude refactor` để cải thiện mã. Ví dụ: `claude refactor --language python --input mycode.py`.

Đánh giá: Hiệu quả khi cải thiện mã, nhưng nên xem xét kiến trúc tổng thể.

BÀI HỌC AI HÔM NAY CHO DEV

1. Tối ưu chi phí & hiệu năng LLM

2. Để phát triển ứng dụng AI hiệu quả, việc tối ưu chi phí và hiệu năng của Model Language lớn (LLM) là rất quan trọng. Điều này giúp giảm thiểu chi phí vận hành và cải thiện hiệu suất của ứng dụng. Dev cần biết cách tối ưu hóa LLM để đảm bảo ứng dụng hoạt động mượt mà và tiết kiệm chi phí.

3. Ví dụ, có thể sử dụng kỹ thuật fine-tuning và LoRA (Low-Rank Adaptation) để tối ưu hóa LLM cho use case cụ thể, giúp giảm thiểu số lượng tham số và cải thiện hiệu suất.

4. Tip: Để bắt đầu, hãy xem xét sử dụng các thư viện như Hugging Face Transformers hoặc TensorFlow để thực hiện fine-tuning và LoRA cho LLM của bạn, và theo dõi các metric như latency và throughput để đánh giá hiệu suất.

Luôn đi đầu trong thế giới AI! · Stay ahead in AI!

Nguồn: Google News · Groq AI