

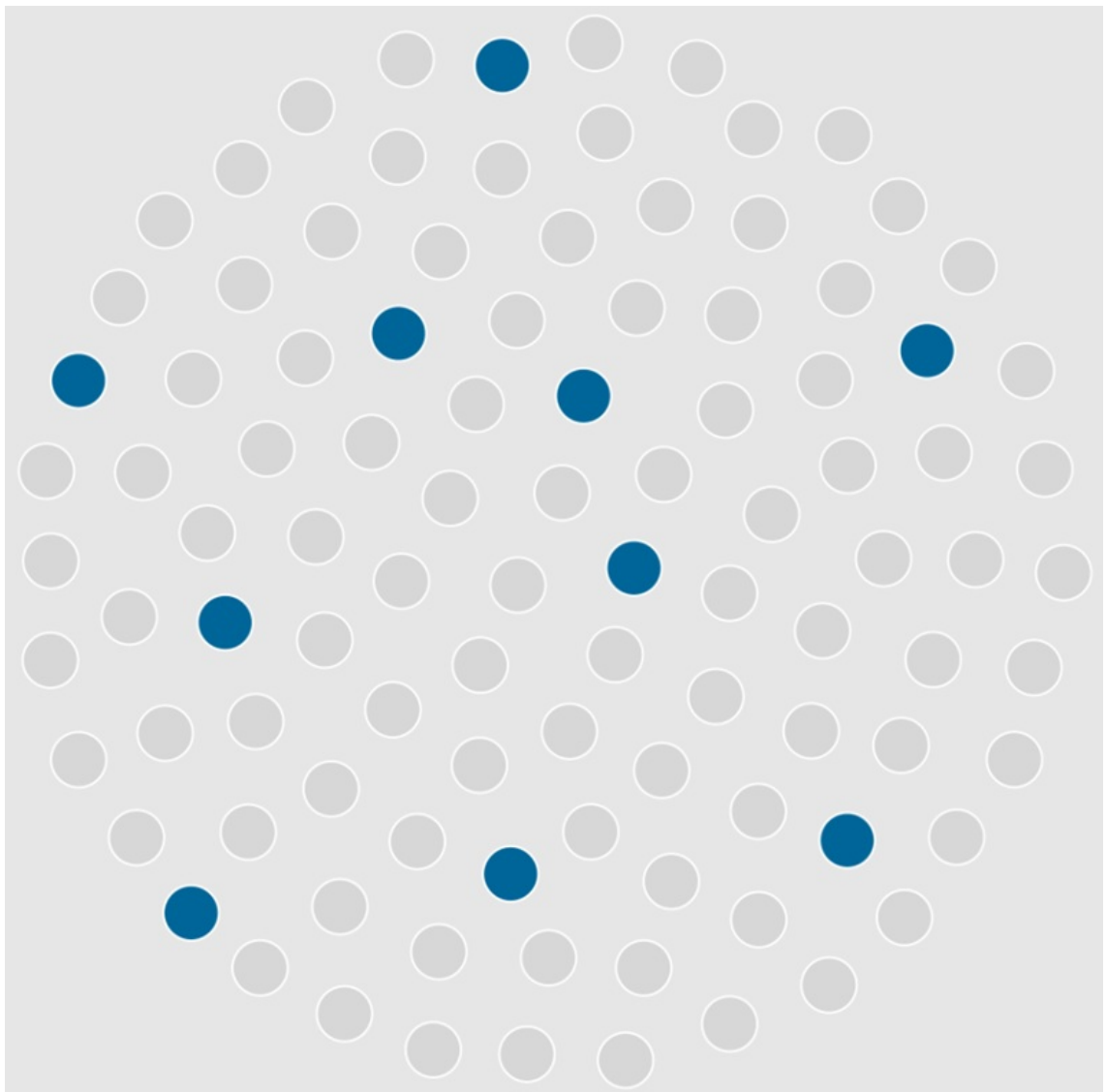
Internet có bao nhiêu nội dung được viết bằng AI?

ISSN: 2734-9195 13:21 21/08/2026

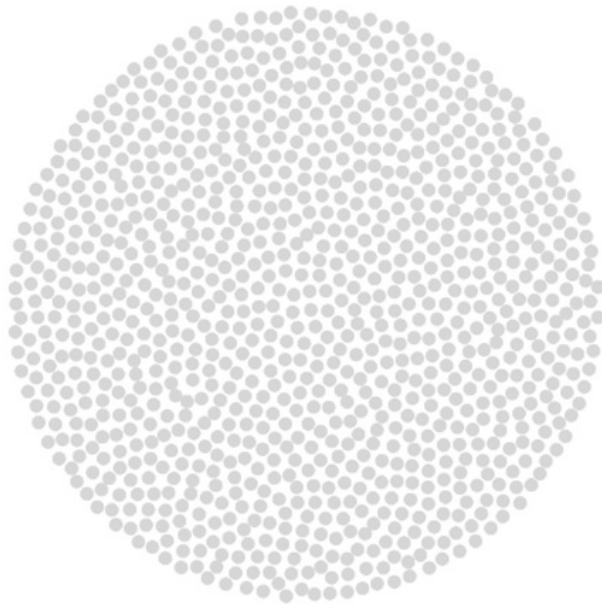
Internet từng được kỳ vọng là không gian mở rộng tri thức của nhân loại. AI đang khiến không gian ấy có khả năng sản xuất nội dung với tốc độ chưa từng có. Nhưng nhiều chữ hơn không đồng nghĩa với nhiều trí tuệ hơn.

Tháng 11/2022, OpenAI lần đầu tiên đưa ChatGPT đến với công chúng. Chưa đầy 4 năm sau, khoảng một nửa người trưởng thành tại Mỹ cho biết họ sử dụng các chatbot được vận hành bằng **trí tuệ nhân tạo (AI)**, trong đó 24% nói rằng họ sử dụng các công cụ này hằng ngày.

Khả năng tạo ra những văn bản có ngôn ngữ gần giống cách con người viết của các công cụ AI đã đặt ra một câu hỏi căn bản về Internet hiện đại: Có bao nhiêu nội dung trên mạng hiện nay được viết bởi AI, thay vì do con người tạo ra?



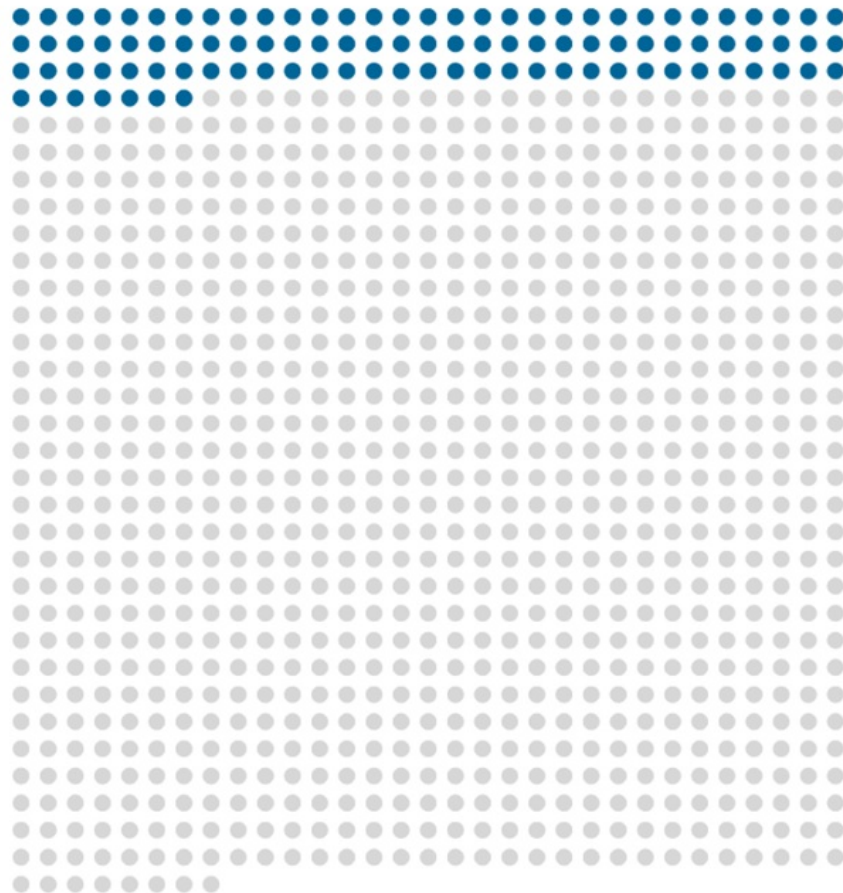
Để tìm hiểu câu hỏi này, chúng tôi sử dụng Common Crawl - kho lưu trữ dữ liệu web quy mô lớn để thu thập gần nửa triệu trang web bằng tiếng Anh trong vòng 5 năm trở lại đây, bắt đầu từ khoảng 2 năm trước khi ChatGPT ra mắt. Sau đó, chúng tôi đưa phần văn bản của các trang này qua một công cụ phát hiện AI có tên Open Pangram, nhằm xác định có bao nhiêu trang có khả năng được viết hoặc được AI biên tập ở mức độ đáng kể.



Các chấm trên biểu đồ đại diện cho một mẫu ngẫu nhiên gồm 10.000 trang web được thu thập vào tháng 07/2026.

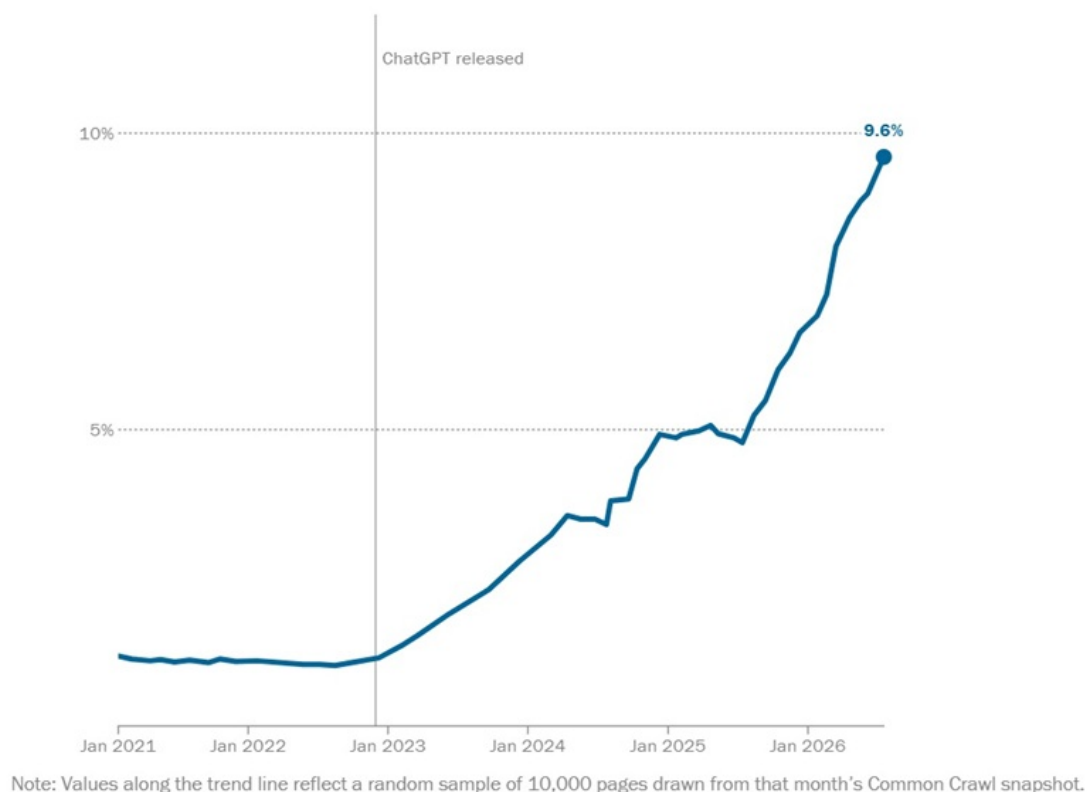
Cùng với nhau, chúng tạo nên một lát cắt cho thấy Internet tiếng Anh trông như thế nào tại thời điểm đó.

Sử dụng một mô hình học máy (machine learning), chúng tôi kiểm tra văn bản trên từng trang để tìm kiếm những dấu hiệu cho thấy nội dung có khả năng do AI tạo ra. Mô hình này phân tích các mẫu hình ngôn ngữ, nhận diện những từ ngữ, cụm từ và đặc điểm ngôn ngữ có xu hướng xuất hiện ở văn bản do AI tạo ra nhiều hơn so với văn bản do con người viết.



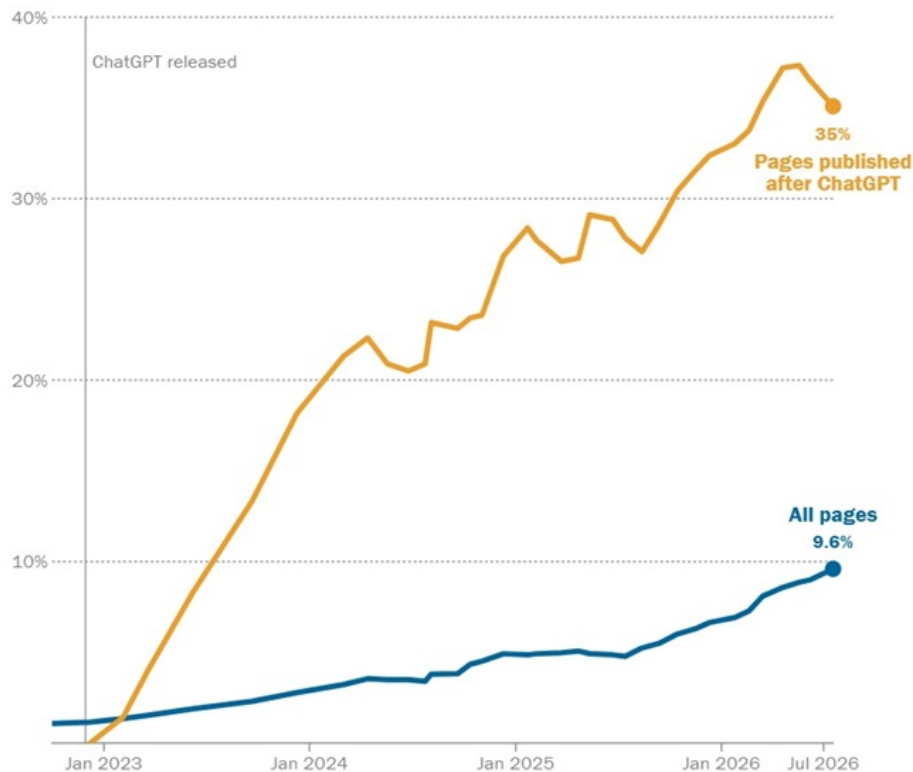
Trong toàn bộ mẫu khảo sát, 10% số trang cho thấy những dấu hiệu đáng kể của việc sử dụng AI trong quá trình tạo lập nội dung.

Đây là một phần của xu hướng gia tăng bắt đầu từ cuối năm 2022, khi **ChatGPT** lần đầu được phát hành, và tiếp tục diễn ra khi những chatbot AI khác như Claude và Gemini xuất hiện. Theo thời gian, tỷ lệ các trang web cho thấy dấu hiệu có sự tham gia của AI ngày càng tăng.



Con số một trên mười trang web tính đến tháng 07/2026 thoát nhìn có thể không quá lớn. Nhưng Internet là sự pha trộn giữa những nội dung mới và cũ. Nhiều trang trong các mẫu ngẫu nhiên này rõ ràng không thể được viết bởi AI, đơn giản vì chúng đã xuất hiện từ trước khi công nghệ này trở nên phổ biến.

Nếu loại các trang web cũ khỏi mẫu khảo sát và chỉ xem xét những trang được xuất bản sau khi ChatGPT ra mắt, xu hướng trở nên rõ rệt hơn nhiều.



Note: Values along the trend line reflect a random sample of 10,000 pages drawn from that month's Common Crawl snapshot.

Trong lát cắt Internet vào tháng 07/2026, dấu hiệu cho thấy AI tham gia vào quá trình tạo lập nội dung xuất hiện ở hơn một phần ba số trang được xuất bản sau khi ChatGPT ra mắt. Kết quả này tương đồng với những nghiên cứu khác cho thấy một tỷ lệ lớn các trang web được xuất bản gần đây có khả năng đã được AI viết hoặc biên tập ở mức độ đáng kể.

Nội dung do AI tạo ra phổ biến nhất ở đâu trên Internet?

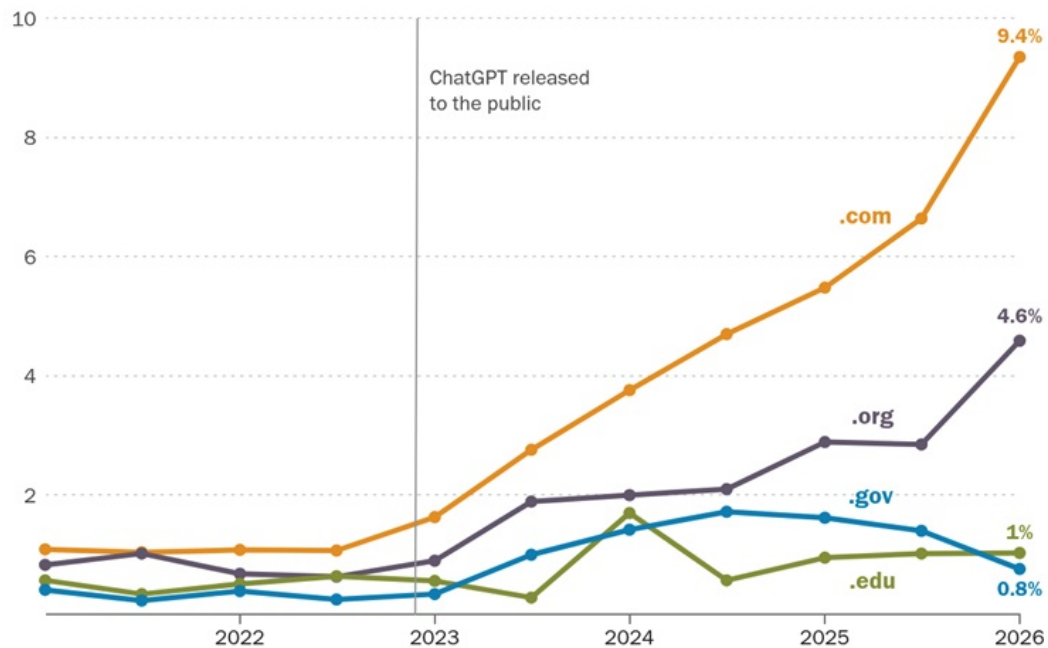
Nội dung do AI tạo ra không phân bố đồng đều trên toàn bộ Internet.

Khi ChatGPT lần đầu xuất hiện, những kiểu mẫu ngôn ngữ có thể là dấu hiệu của việc sử dụng AI xuất hiện với tỷ lệ tương đối giống nhau trên các tên miền cấp cao phổ biến (**top-level domains**) như .com, .org, .edu và .gov.

Nhưng trong các mẫu khảo sát năm 2026, khoảng một trên mười trang thuộc **tên miền .com** cho thấy dấu hiệu của nội dung do AI tạo ra. Tỷ lệ này cao khoảng gấp đôi so với các trang thuộc tên miền .org ở mức **4,6%** và cao gấp khoảng 10 lần so với các trang thuộc .edu hoặc .gov, vốn đều ở khoảng **1%**.

Signs of AI authorship are less common in .edu and .gov domains

% of webpages showing significant signs of AI editing or authorship



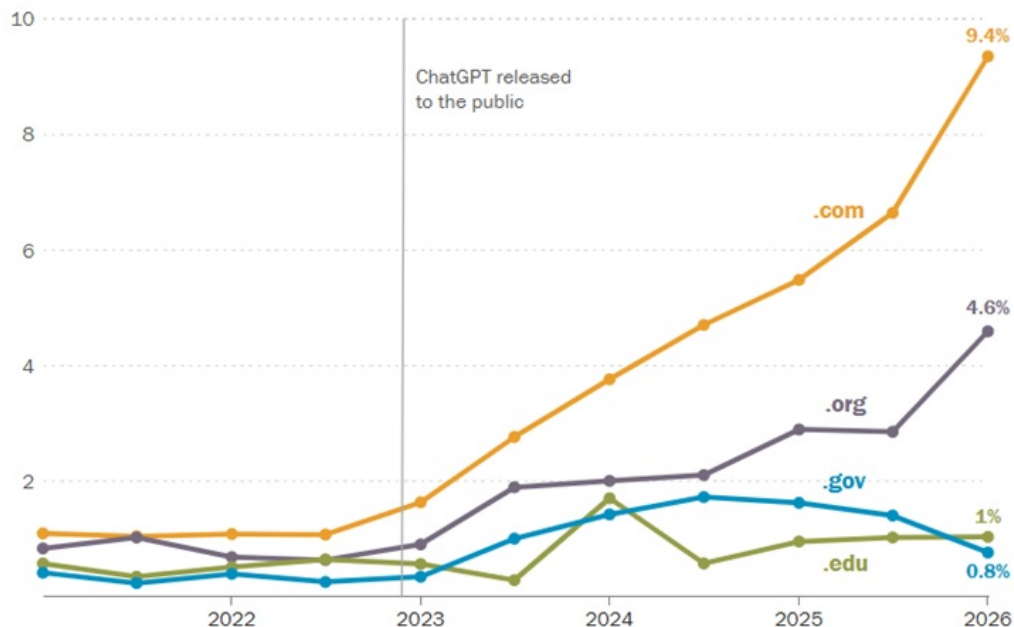
Note: Plotted points represent six-month averages.

Source: Pew Research Center analysis of 490,000 English-language webpage texts sampled from the Common Crawl web repository and analyzed using Open Pangram.

PEW RESEARCH CENTER

Dấu hiệu cho thấy AI là tác giả ít phổ biến hơn ở các tên miền .edu và .gov.

Tỷ lệ phần trăm các trang web cho thấy dấu hiệu rõ ràng của việc chỉnh sửa hoặc viết bởi AI.



Lưu ý: Các điểm trên biểu đồ thể hiện giá trị trung bình trong sáu tháng.

Nguồn: Phân tích của Trung tâm Nghiên cứu Pew dựa trên 490.000 đoạn văn bản trang web tiếng Anh được lấy mẫu từ kho lưu trữ web Common Crawl và được phân tích bằng Open Pangram.

TRUNG TÂM NGHIÊN CỨU PEW

Điều đó cho thấy sự hiện diện của AI trên Internet không chỉ đang gia tăng về quy mô mà còn có sự khác biệt đáng kể giữa các loại không gian trực tuyến.

Những đặc điểm phổ biến của văn bản do AI tạo ra là gì?

Các mô hình phát hiện AI không hoàn hảo đôi khi chúng có thể nhận diện nhầm một văn bản do con người viết là có dấu hiệu của AI, và ngược lại.

Tuy nhiên, khi xem xét những tập hợp văn bản rất lớn, chúng ta bắt đầu nhận ra rằng một số kiểu dấu câu, từ ngữ và cụm từ xuất hiện với tần suất cao hơn đáng kể trong nội dung do AI tạo ra so với văn bản do con người viết.

[Được viết bởi AI

Internet đã phát triển vượt xa một mạng lưới đơn giản gồm những trang được kết nối với nhau, đó là một hệ sinh thái sống động và không ngừng biến đổi, được định hình bởi các thuật toán, người sáng tạo và công chúng, đồng thời được thúc đẩy bởi những làn sóng nội dung do AI tạo ra, làm mờ ranh giới giữa cách biểu đạt của con người và máy móc. Khi người dùng tiếp tục đi sâu vào những dòng bài đăng, video và giọng nói tổng hợp bất tận, một câu hỏi mang tính bước ngoặt xuất hiện về tính xác thực, quyền sở hữu và niềm tin bởi vấn đề không chỉ là thông tin, mà còn là ảnh hưởng.

Dấu gạch ngang dài (em dash —)

Dấu phẩy Oxford (Oxford comma)

Từ vựng đặc trưng của AI

Phép song song phủ định (negative parallelism)]

Các mô hình AI được huấn luyện trên những tập dữ liệu khổng lồ gồm văn bản do con người viết, từ đó học cách mô phỏng những mẫu hình và phong cách của con người.

Đôi khi, một số kiểu văn bản nhất định xuất hiện quá nhiều trong dữ liệu huấn luyện. Vì vậy, các mô hình được huấn luyện trên những dữ liệu ấy có thể sử dụng một số từ, cụm từ hoặc đặc điểm ngôn ngữ với tần suất cao hơn mức con người thường sử dụng.

Chẳng hạn, nhiều văn bản báo chí và học thuật sử dụng **dấu gạch ngang dài (em dash —)**, một dấu câu dùng để thể hiện sự ngắt quãng trong ý tưởng hoặc tách một thành phần chú thích khỏi câu. Các mô hình AI có xu hướng sử dụng loại dấu gạch ngang này nhiều hơn đáng kể so với con người.

AI cũng có xu hướng liệt kê sự vật thành ba thành phần, đồng thời sử dụng dấu phẩy Oxford (Oxford comma) trong các danh sách nhiều thành phần thường xuyên hơn tác giả là con người.

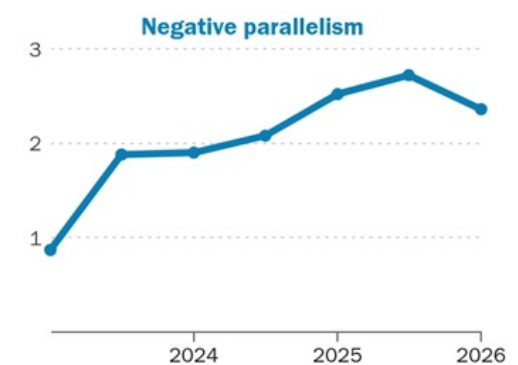
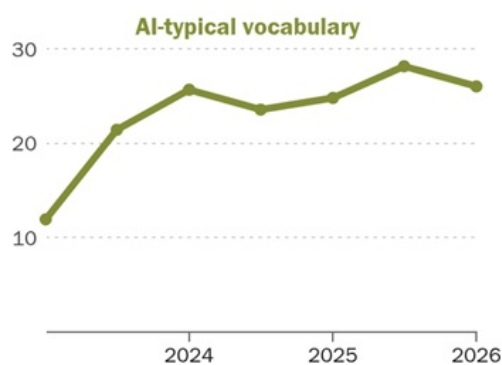
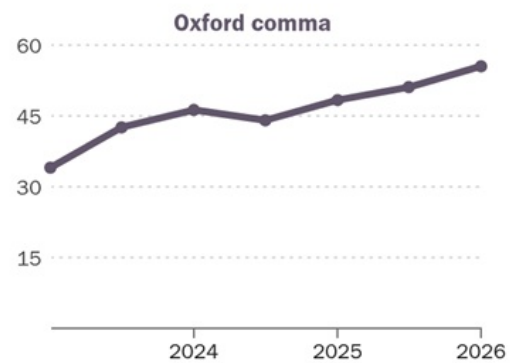
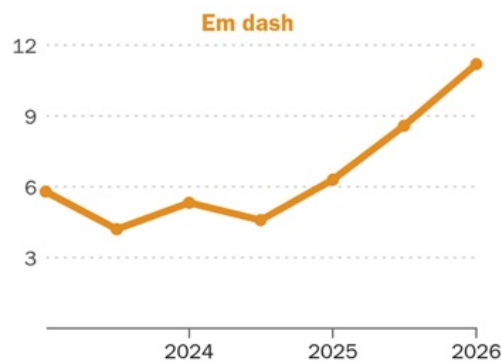
Khi ngày càng có nhiều nội dung do AI tạo ra xuất hiện trên Internet, những đặc điểm này cùng nhiều “dấu vết” khác của AI (**AI tells**) cũng ngày càng trở nên phổ biến trên các trang web.

Nếu so sánh Internet hiện nay với một lát cắt được ghi nhận vào năm 2023:

- + Dấu gạch ngang dài (em dash) xuất hiện với tần suất cao hơn khoảng hai lần.
- + Dấu phẩy Oxford tăng 63%.
- + Một số từ mà các mô hình AI đặc biệt ưa dùng — chẳng hạn *delve* (đào sâu), *interplay* (sự tương tác qua lại) hoặc *testament* (minh chứng) — đã tăng hơn gấp đôi về tần suất sử dụng.
- + Việc sử dụng “phép song song phủ định” (negative parallelism) để cấu trúc một phép so sánh — chẳng hạn “không chỉ là X, mà còn là Y” — đã tăng gần gấp ba lần, mặc dù nhìn chung kiểu diễn đạt này vẫn tương đối hiếm.

Key words, phrases and punctuation favored by AI have become more common on the web

Uses of ___ per 10,000 words

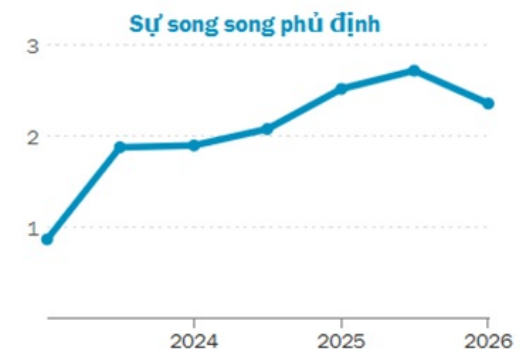
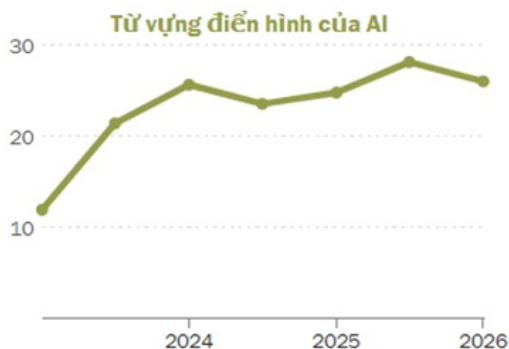
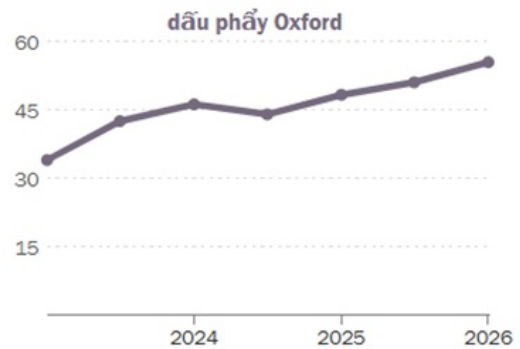
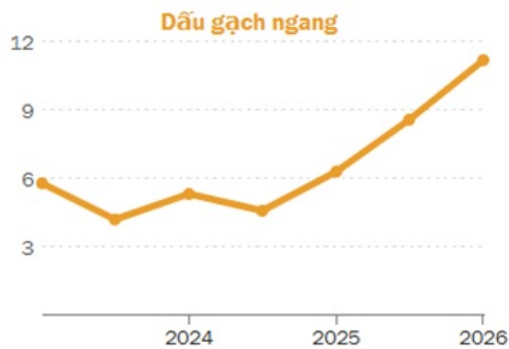


Note: AI-typical vocabulary includes “additionally,” “align with,” “boasts,” “bolstered,” “crucial,” “delve,” “emphasizing,” “enduring,” “enhance,” “essential,” “fostering,” “garner,” “highlight,” “interplay,” “intricate,” “key,” “landscape,” “meticulous,” “perfectly,” “pivotal,” “showcase,” “significant,” “tapestry,” “testament,” “underscore,” “valuable,” “vibrant.” Samples filtered to pages published after Nov. 30, 2022. Plotted points represent six-month averages.
Source: Pew Research Center analysis of 490,000 English-language webpage texts sampled from the Common Crawl web repository.

PEW RESEARCH CENTER

Các từ khóa, cụm từ và dấu câu được trí tuệ nhân tạo ưa chuộng ngày càng trở nên phổ biến trên web.

Số lần sử dụng ___ trên 10.000 từ



Lưu ý: Từ vựng điển hình của AI bao gồm "thêm vào đó," "phù hợp với," "khoe khoang," "củng cố," "quan trọng," "đi sâu," "nhấn mạnh," "bền bỉ," "nâng cao," "thiết yếu," "thúc đẩy," "thu thập," "làm nổi bật," "tương tác," "phức tạp," "chủ chốt," "cảnh quan," "tỉ mỉ," "hoàn hảo," "quan trọng," "trưng bày," "đáng kể," "di chúc," "gạch chân," "có giá trị," "sôi động." Các mẫu được lọc theo các trang được xuất bản sau ngày 30 tháng 11 năm 2022. Các điểm được vẽ trên biểu đồ thể hiện mức trung bình trong sáu tháng.

Nguồn: Phân tích của Trung tâm Nghiên cứu Pew dựa trên 490.000 văn bản trang web tiếng Anh được lấy mẫu từ kho dữ liệu web Common Crawl.

TRUNG TÂM NGHIÊN CỨU PEW

Nguyên mẫu biểu đồ tiếng Anh và bản được Việt hóa

Tất nhiên, chỉ riêng việc một văn bản sử dụng dấu gạch ngang dài hay dấu phẩy Oxford không có nghĩa chắc chắn rằng văn bản đó được tạo ra bằng AI. Con người cũng sử dụng những cách viết và dấu câu này.

Nhưng khi xem xét một số lượng lớn tài liệu cùng lúc, chúng ta có thể nhận ra những khuôn mẫu cho thấy trung bình tác giả là con người và các mô hình AI thường lựa chọn những đặc điểm ngôn ngữ này với tần suất như thế nào.

Hơn nữa, những mô hình phát hiện AI tinh vi như mô hình được sử dụng trong phân tích này không chỉ đơn giản tìm kiếm một vài "dấu hiệu nhận biết" riêng lẻ. Chúng còn học cách nhận diện những mẫu hình thống kê tinh vi hơn trong việc lựa chọn từ ngữ và cấu trúc câu, qua đó giúp xác định liệu một văn bản cụ thể có khả năng được tạo ra bởi AI hay không.

Góc nhìn Tạp chí:

Khi AI viết ngày càng nhiều, điều gì còn làm nên giá trị của một văn bản?

Những con số trong nghiên cứu này đáng chú ý không chỉ vì chúng cho thấy AI đang hiện diện ngày càng sâu trên Internet. Đằng sau câu chuyện về 10% hay hơn một phần ba số trang web mới có dấu hiệu AI tham gia còn là một câu hỏi rộng hơn: Nếu máy móc ngày càng có khả năng tạo ra những câu chữ giống con người, điều gì làm nên giá trị đích thực của một lời viết?



(Ảnh: Internet)

Từ góc nhìn Phật giáo, có lẽ câu trả lời không nằm trước hết ở việc **ai viết**, con người hay máy móc mà ở **tâm ý, mục đích và hệ quả** của lời viết ấy.

Trong giáo lý Phật giáo, **chính ngữ (sammā-vācā)** không đơn thuần là nói đúng ngữ pháp hay diễn đạt trôi chảy. Chánh ngữ gắn với việc tránh lời nói dối, lời nói gây chia rẽ, lời nói thô ác và lời nói phù phiếm; đồng thời hướng lời nói đến sự chân thật, hòa hợp, có ích và đúng thời.

Đặt nguyên tắc ấy vào thời đại AI, vấn đề trở nên đặc biệt đáng suy ngẫm.

Một văn bản do AI tạo ra có thể mạch lạc, đẹp đẽ, giàu thông tin và thậm chí rất “giống người”. Nhưng sự trôi chảy của câu chữ không đồng nghĩa với sự chân thật của nội dung. Một bài viết có thể hoàn hảo về ngữ pháp nhưng vẫn chứa thông tin sai; có thể giàu cảm xúc nhưng lại được tạo ra nhằm thao túng cảm xúc; có thể mang dáng vẻ học thuật nhưng thiếu kiểm chứng; có thể nói về lòng

từ bí nhưng được sản xuất hàng loạt chỉ để thu hút lượt xem.

Ở đây xuất hiện một nghịch lý của thời đại AI: **càng khó phân biệt người viết với máy viết, trách nhiệm của người sử dụng AI càng lớn.**

AI có thể tạo câu chữ, nhưng chưa thể thay thế trách nhiệm đạo đức của người đưa câu chữ ấy vào đời sống.

Một người dùng AI để hỗ trợ nghiên cứu, dịch thuật, kiểm tra ngôn ngữ hay mở rộng khả năng tiếp cận tri thức là một chuyện. Một người sử dụng AI để tạo ra hàng nghìn bài viết thiếu kiểm chứng, làm loãng không gian thông tin, đánh lừa độc giả hoặc thao túng nhận thức lại là chuyện khác.

Vì vậy, câu hỏi quan trọng có lẽ không chỉ là: “Bao nhiêu phần trăm bài viết/nội dung trên Internet được viết bằng AI?”

Mà cần hỏi thêm: “Bao nhiêu phần trăm trong số đó được tạo ra trong sự tỉnh thức và trách nhiệm?”

Đây cũng là nơi khái niệm chính niệm (sati) có thể mở ra cách tiếp cận đáng suy nghĩ đối với AI.

Chính niệm không có nghĩa là từ chối công nghệ. Ngược lại, chính niệm là khả năng nhận biết rõ mình đang làm gì, vì sao mình làm điều đó và hành động ấy có thể dẫn đến đâu.

Một người có thể dùng AI rất nhiều mà vẫn tỉnh thức. Và cũng có thể không dùng AI nhưng vẫn tạo ra những thông tin thiếu trách nhiệm.

Do đó, vấn đề không nằm ở bản thân công cụ. *Vấn đề nằm ở cách chúng ta sử dụng công cụ và trách nhiệm mà chúng ta giữ lại sau khi công cụ đã hoàn thành phần việc được chia sẻ, giao phó.*

Trong bối cảnh ấy, việc phát hiện văn bản do AI tạo ra chỉ là một nửa câu chuyện. Nửa còn lại và có lẽ quan trọng hơn là xây dựng một văn hóa thông tin trong đó con người biết kiểm chứng, biết hoài nghi đúng lúc, biết chịu trách nhiệm với điều mình xuất bản và biết dừng lại trước khi biến tốc độ của máy thành tốc độ của sự thiếu tỉnh thức.

Internet từng được kỳ vọng là không gian mở rộng tri thức của nhân loại. AI đang khiến không gian ấy có khả năng sản xuất nội dung với tốc độ chưa từng có.

Nhưng **hiều chữ hơn không đồng nghĩa với nhiều trí tuệ hơn.**

Thay lời kết

AI đang học cách viết giống con người.

Điều đáng suy nghĩ là trong khi máy móc ngày càng học được cách tạo ra những câu chữ giống chúng ta, chúng ta có đang ngày càng biết rõ mình muốn nói gì hay không?

Môi trường Internet tràn ngập văn bản do AI tạo ra chưa chắc đã nghèo nàn. AI có thể giúp con người dịch chuyển tri thức nhanh hơn, tiếp cận thông tin rộng hơn và mở ra những khả năng sáng tạo mới.

Nhưng tốc độ cũng có những mặt trái.

Khi một ý tưởng có thể biến thành hàng trăm bài viết chỉ trong vài phút, sự khan hiếm đáng lo ngại nhất có thể không còn là **nội dung**, mà là **sự chú tâm**. Khi câu chữ trở nên rẻ và vô hạn, thứ trở nên quý giá hơn lại là suy nghĩ được cân nhắc cẩn trọng, là thông tin được kiểm chứng và một lời nói xuất phát từ trách nhiệm.

Phật giáo từ lâu đã đặt vấn đề về nghiệp của lời nói: lời nói không chỉ được đánh giá bởi âm thanh hay hình thức, mà còn bởi ý hướng và hệ quả tạo tác.

Trong kỷ nguyên AI, nguyên tắc ấy có thể được đọc lại theo một cách mới: AI có thể tạo ra câu chữ. Nhưng chúng ta vẫn phải chịu trách nhiệm về điều mình cho phép câu chữ ấy làm với thế giới.

Vì thế, có lẽ câu hỏi lớn nhất của Internet trong thời đại AI không phải là: “Bao nhiêu phần trăm nội dung được viết bởi AI?”

Mà là: “Trong bao nhiêu phần trăm nội dung ấy, chúng ta vẫn còn hiện diện với sự tỉnh thức?”

Lời cảm ơn

Phân tích này được thực hiện bởi **Samuel Bestvater**, nhà khoa học dữ liệu cao cấp (*senior data scientist*), đồng thời là tác giả của bài viết.

Carson TerBush, chuyên viên thiết kế đồ họa thông tin (*associate information graphics designer*), thực hiện phần đồ họa và hoạt ảnh, đồng thời đóng góp vào quá trình phát triển phiên bản web, với sự hỗ trợ của **Chris Baranovski**, kỹ sư trưởng phụ trách nội dung biên tập (*lead engineer - editorial content*).

Janakee Chavda, nhà sản xuất nội dung số (*associate digital producer*), phụ trách sản xuất báo cáo trên nền tảng web.

Mã nguồn phục vụ phân tích được **Skyler Seets**, trợ lý khoa học xã hội tính toán (*computational social science assistant*), rà soát; bài viết được **Anna Lieb**, chuyên viên phân tích khoa học xã hội tính toán (*computational social science analyst*), kiểm tra.

Phân biên tập ngôn ngữ (*copy editing*) do **Anna Jackson**, chuyên viên biên tập (*editorial specialist*), thực hiện.

Aaron Smith, Giám đốc nghiên cứu **Data Labs**, cung cấp định hướng biên tập cho dự án.

Các tác giả: **Samuel Bestvater, Aaron Smith, Carson TerBush, Chris Baronavski**

và Janakee Chavda

Các tác giả: **Samuel Bestvater, Aaron Smith, Carson TerBush, Chris Baronavski và Janakee Chavda**

Chuyển ngữ và biên tập: **Thường Nguyễn**

Nguồn: **<https://www.pewresearch.org/data-labs/2026/08/20/how-much-of-the-internet-is-written-with-ai/>**