

Sử dụng tác phẩm được cấp phép CC để đào tạo AI

Dịch sang tiếng Việt: Lê Trung Nghĩa

Dịch xong: 28/06/2025

Bản gốc tiếng Anh: <https://creativecommons.org/wp-content/uploads/2025/05/Using-CC-licensed-Works-for-AI-Training.pdf>

Using CC-licensed Works for AI Training



Sử dụng tác phẩm được cấp phép CC để đào tạo AI. Bản quyền © 2025 của Creative Commons được cấp giấy phép [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)



Using CC-licensed Works for AI Training © 2025 by Creative Commons is licensed under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)

Sử dụng tác phẩm được cấp phép CC để đào tạo AI

Việc áp dụng luật bản quyền vào đào tạo AI rất phức tạp. Trên toàn thế giới, có những ngoại lệ và hạn chế đối với bản quyền tạo điều kiện cho đào tạo AI, nhưng luật pháp có sự khác biệt đáng kể. Phân tích theo luật quốc nội cũng thường dựa trên thực tế.

Vì giấy phép CC là giấy phép bản quyền, nên việc áp dụng giấy phép CC vào đào tạo AI cũng phức tạp tương tự. **Tóm lại, có nhiều trường hợp không yêu cầu tuân thủ các điều kiện của giấy phép CC khi sử dụng các tác phẩm được cấp phép CC để đào tạo AI.** Để biết phiên bản phân tích dài, hãy xem phần “Tôi có phải tuân thủ giấy phép CC đối với dữ liệu đào tạo của mình không?” ở phần sau của tài liệu này.

Do tính phức tạp này, một số nhà phát triển và nhà nghiên cứu AI có thể chọn tuân thủ giấy phép CC đối với dữ liệu đào tạo **trong mọi trường hợp**, để giảm thiểu rủi ro pháp lý hoặc như một cử chỉ thiện chí, hoặc cả hai. Hướng dẫn sau đây được thiết kế cho các trường hợp đó và không nhằm mục đích đưa ra quan điểm về việc bản quyền có được áp dụng hay không và khi nào.

Sử dụng các thành phần của giấy phép CC làm hướng dẫn cho đào tạo AI

Điều đáng lưu ý là việc tuân thủ hướng dẫn này gần như chắc chắn sẽ dẫn đến việc tuân thủ quá mức theo luật bản quyền và bản thân các giấy phép Creative Commons. Nó giả định phiên bản hạn chế nhất của các sự kiện và luật để đưa ra cách tiếp cận bảo thủ nhất.

Ghi công (Attribution): Tất cả các giấy phép CC đều yêu cầu ghi công để xác định người tạo ra tài liệu được cấp phép. Khi nói đến đào tạo mô hình nói chung, ghi công có thể là một liên kết đơn giản đến nguồn của tập dữ liệu được sử dụng để đào tạo mô hình. Ví dụ: "Dữ liệu đào tạo từ tập dữ liệu LAION [đường liên kết]."

Ngoài ra, việc ánh xạ đầu ra mô hình vào một tác phẩm cụ thể trong dữ liệu đào tạo hiện chỉ có thể thực hiện được trong những trường hợp hạn chế, chẳng hạn như các trường hợp sử dụng cụ thể của việc tạo sinh được tăng cường truy xuất hoặc RAG (Retrieval-Augmented Generation). Khi RAG hoặc các phương pháp khác khả dụng, việc ghi công cho tác phẩm được cấp phép CC gắn với đầu ra mô hình cụ thể bằng một đường liên kết đến nguồn là lý tưởng.

Chia sẻ tương tự (ShareAlike): Hai trong số các giấy phép CC yêu cầu các bản chuyển thể phải được cung cấp theo cùng một giấy phép. Khi dữ liệu đào tạo tuân theo điều

kiện Chia sẻ tương tự, đầu ra của mô hình và chính mô hình, nếu được chia sẻ công khai, phải được cung cấp theo cùng giấy phép CC như tác phẩm gốc khi áp dụng phương pháp bảo thủ này.

Phi thương mại (NonCommercial): Hai trong số các giấy phép CC chỉ cấp phép cho mục đích sử dụng Phi thương mại. Khi dữ liệu đào tạo phải tuân theo hạn chế Phi thương mại, việc tuân thủ giấy phép CC yêu cầu mục đích sử dụng của bạn không được "chủ yếu nhằm mục đích hoặc hướng đến lợi thế thương mại hoặc bồi thường tiền tệ". Nói cách khác, việc sao chép tác phẩm trong quá trình đào tạo mô hình, cũng như việc sử dụng và phân phối mô hình đã đào tạo sau đó, sẽ cần phải nhằm mục đích phi thương mại.

Không phái sinh (NoDerivatives): Hai trong số các giấy phép CC cấm việc tạo ra các tác phẩm phái sinh. Mặc dù trong nhiều trường hợp, cả mô hình AI và đầu ra của nó đều không được coi là tác phẩm phái sinh của dữ liệu đào tạo theo luật bản quyền, nhưng tác phẩm được cấp phép CC phải tuân theo hạn chế Không phái sinh không nên được sử dụng làm dữ liệu đào tạo nếu áp dụng cách tiếp cận bảo thủ này.

Tôi có phải tuân thủ giấy phép CC đối với dữ liệu đào tạo của mình không?

Tùy thuộc. Giấy phép CC áp dụng cho việc sử dụng tại trong mọi tình huống yêu cầu phải có sự cho phép theo luật bản quyền. Chúng không có hiệu lực đối với các tình huống không yêu cầu phải có sự cho phép theo luật bản quyền, chẳng hạn như khi áp dụng ngoại lệ bản quyền. Việc áp dụng luật bản quyền vào đào tạo AI rất phức tạp, tùy thuộc vào khu vực pháp lý nơi sử dụng và các vụ kiện tụng liên quan đến đào tạo AI tạo ra vẫn đang diễn ra.

Có hai khía cạnh chính đối với phân tích bản quyền khi nói đến việc sử dụng các tác phẩm có bản quyền làm dữ liệu đầu vào cho đào tạo: đào tạo và ghi nhớ.¹

Thu thập và chuẩn bị bộ dữ liệu để đào tạo: Quá trình đào tạo các mô hình AI gần như luôn yêu cầu sao chép các tác phẩm được sử dụng làm dữ liệu đào tạo. Khi các tác phẩm đó có bản quyền, bước đào tạo có thể có ý nghĩa về bản quyền, mặc dù nó khác nhau tùy thuộc vào luật hiện hành. Ở một số khu vực pháp lý, chẳng hạn như Nhật Bản, việc sao chép bắt buộc để đào tạo nằm trong ngoại lệ và giới hạn của bản quyền và không yêu cầu phải có sự cho phép. Ở những nơi khác, chẳng hạn như EU, các hành vi này nằm trong ngoại lệ và giới hạn của luật bản quyền, ngoại trừ trong trường hợp

chủ sở hữu quyền của tác phẩm đã từ chối rõ ràng. Ở những nơi khác, chẳng hạn như Úc hoặc Vương quốc Anh, việc đào tạo có thể được thực hiện theo những ngoại lệ hiện hành, hạn chế (ví dụ: Vương quốc Anh có một ngoại lệ hạn chế liên quan đến khai thác văn bản và dữ liệu cho mục đích nghiên cứu khoa học phi thương mại), nhưng ngoài ra, việc sao chép cần thiết cho mục đích đào tạo thường bị hạn chế bởi bản quyền.

Một số đầu ra có những vấn đề riêng biệt, chẳng hạn như khi người dùng nhắc mô hình hình ảnh tạo ra một ký tự có bản quyền. Nếu ký tự có bản quyền đó được cấp phép theo CC, thì việc sử dụng sẽ phải được chứng minh dựa trên việc tuân thủ giấy phép hoặc các ngoại lệ và hạn chế của bản quyền. Các câu hỏi về việc ai sẽ chịu trách nhiệm (người dùng, nhà cung cấp mô hình hoặc cả hai) về hành vi vi phạm nằm ngoài phạm vi của Câu hỏi thường gặp này.

Ghi nhớ trong quá trình đào tạo: Có những trường hợp mà bản thân các mô hình AI sao chép và lưu trữ một số biểu thức có thể có bản quyền của các tác phẩm mà chúng được đào tạo. Điều này thường được gọi là ghi nhớ và đây là cách chính khác mà bản quyền có thể bị ảnh hưởng bởi quá trình đào tạo mô hình AI. Hiện tại, cách duy nhất để biết liệu một mô hình có ghi nhớ nội dung hay không là xác định các đầu ra của mô hình về cơ bản giống với bản gốc. Điều này có nghĩa là không thể định lượng được mức độ ghi nhớ trong bản tóm tắt. Điều quan trọng nữa là phải lưu ý rằng không phải mọi trường hợp ghi nhớ đều là vi phạm. Ở nhiều mức độ khác nhau, các nhà phát triển mô hình sẽ nỗ lực điều chỉnh phương pháp của mình để tránh phải ghi nhớ và điều này có thể làm giảm nhưng không thể loại bỏ hoàn toàn nguy cơ vi phạm bản quyền.

Với điều này trong tâm trí, câu hỏi tiếp theo trong phân tích là khi nào các điều kiện cấp phép cụ thể có hiệu lực. Điều kiện ghi công (BY), điều kiện Chia sẻ tương tự (SA) và hạn chế Không phái sinh (ND) đều chỉ được kích hoạt khi chia sẻ công khai tác phẩm gốc hoặc chuyển thể tác phẩm gốc, tùy trường hợp. Điều này có nghĩa là, ví dụ, luật bản quyền sẽ yêu cầu ghi công khi mô hình ghi nhớ và do đó lưu trữ một bản sao của biểu thức về cơ bản giống với tác phẩm được cấp phép theo CC mà mô hình được đào tạo và không có ngoại lệ nào được áp dụng, vì việc chia sẻ mô hình hoặc chia sẻ các đầu ra về cơ bản giống nhau sẽ cấu thành việc chia sẻ tác phẩm. Ngược lại, hạn chế Phi thương mại (NC) áp dụng cho tất cả các mục đích sử dụng yêu cầu phải có sự cho phép theo bản quyền. Xem [sơ đồ luồng này](#) để biết thêm thông tin.