KHUNG NĂNG LỰC SỐ TRONG VIỆC CÁ NHÂN HOÁ TRÍ TUỆ NHÂN TẠO

Hoàng Tuấn Anh[1]

Tóm tắt: *Trong bối cảnh các mô hình trí tuệ nhân tạo (AI) ngày càng đóng vai trò thiết yếu trong đời sống xã hội, nhu cầu cá nhân hóa AI đang trở thành một xu thế không thể đảo ngược. Bài viết này phân tích tiềm năng và tính cấp thiết của việc huấn luyện AI theo cách người dùng mong muốn, đồng thời chỉ ra những năng lực số (digital competencies) mà người dùng và tổ chức cần có để triển khai cá nhân hóa một cách hiệu quả và có đạo đức. Thông qua tổng quan lý thuyết, khảo sát người dùng, thực nghiệm với mô hình LLaMA 3 và GPT-4-turbo và phân tích chi phí-hiệu quả, tác giả chứng minh rằng huấn luyện cá nhân hóa không chỉ khả thi về mặt kỹ thuật mà còn là cơ hội để mở rộng dân chủ hóa AI. Đồng thời bài viết đề xuất khung năng lực số cho cá nhân và tổ chức nhằm tiếp cận AI một cách chủ động, minh bạch và bền vững.*

Từ khóa: *Huấn luyện AI, Năng lực số, trí tuệ nhân tạo (AI).*

Abstract: *As artificial intelligence (AI) models increasingly play a critical role in society, the demand for AI personalization is becoming a dominant and irreversible trend. This paper analyzes the feasibility and necessity of user-directed AI training and identifies the essential digital competencies required by individuals and organizations to implement personalization effectively and ethically. Based on a literature review, user survey, experiments with LLaMA 3 and GPT-4-turbo models, and a cost-effectiveness analysis, the study demonstrates that personalized AI training is both technically viable and a means to enhance AI democratization. The paper proposes a digital competency framework to support proactive, transparent, and sustainable engagement with AI.*

Keywords*: AI training, Digital competencies, artificial intelligence (AI).*

Nhận bài: 04/07/2025 Gửi phản biện: 10/07/2025 Duyệt đăng: 15/08/2025

1. Đặt vấn đề

Trí tuệ nhân tạo (AI) đang phát triển vượt bậc nhờ các mô hình ngôn ngữ lớn (LLMs) như GPT-4-turbo (OpenAI, 2024), Claude 3 (Anthropic, 2024) hay LLaMA 3 (Meta, 2024). Các mô hình này được huấn luyện trên kho dữ liệu khổng lồ và mang lại khả năng sinh ngôn ngữ, suy luận, lập luận và sáng tạo vượt trội. Tuy nhiên, quá trình huấn luyện tập trung và tổng quát hóa đã khiến nhiều mô hình thiếu tính linh hoạt trong việc phản ánh văn hóa, ngữ cảnh hay giá trị cá nhân của người dùng.

1 Cử nhân, Học viện Chính trị - Bộ Quốc phòng

Vấn đề đặt ra là: Liệu AI có thể và nên được huấn luyện theo phong cách, giá trị và mục tiêu riêng biệt của từng người dùng? Câu hỏi này không chỉ mang tính kỹ thuật mà còn là một thách thức đạo đức và xã hội. Hơn thế nữa, để thực hiện điều đó, người dùng cần trang bị những năng lực số cần thiết, bao gồm khả năng xử lý dữ liệu, hiểu biết về AI, kỹ năng quản trị rủi ro và đạo đức công nghệ.

Mục tiêu của bài viết này là xây dựng luận cứ lý thuyết và thực nghiệm cho xu hướng huấn luyện AI theo cách người dùng mong muốn (user-directed training), đồng thời đề xuất khung năng lực số nền tảng cho cá nhân và tổ chức nhằm tiếp cận hiệu quả xu hướng này.

2. Phương pháp nghiên cứu

Bài viết áp dụng phương pháp nghiên cứu hỗn hợp (mixed-methods), kết hợp phân tích định tính và định lượng để đánh giá toàn diện khả năng và nhu cầu cá nhân hóa mô hình AI. Cụ thể:

- *Tổng quan tài liệu:* Khảo cứu các nghiên cứu gần đây về LLMs, kỹ thuật tinh chỉnh mô hình (fine-tuning), năng lực số và cá nhân hóa AI từ các nguồn uy tín như OpenAI, Meta AI, Mozilla Foundation, nhằm xây dựng nền tảng lý luận vững chắc.

- *Phân tích dữ liệu thứ cấp*: Sử dụng dữ liệu khảo sát từ McKinsey (2024) và Mozilla Foundation (2023) để làm rõ xu hướng và kỳ vọng của người dùng trong việc tương tác và tùy biến AI.

- *Thực nghiệm mô hình*: Tinh chỉnh GPT-4-turbo (API) và LLaMA 3 (7B, chạy cục bộ) trên ba loại dữ liệu cá nhân: email, blog và truy vấn tìm kiếm. Kết quả được đánh giá bằng các chỉ số BLEU, ROUGE-L và mức độ tương tác người dùng Engagement Score.

- *Phân tích chi phí – hiệu quả:* Đo lường thời gian, chi phí phần cứng và tài nguyên tính toán để xác định tính khả thi kinh tế của cá nhân hóa ở quy mô nhỏ.

3. Kết quả và bàn luận

3.1 Cơ sở khoa học cho việc hình thành năng lực số trong huấn luyện AI

3.1.1. Các khái niệm nền tảng

Trong bối cảnh chuyển đổi số toàn diện, năng lực số (digital competence) đang trở thành một yêu cầu thiết yếu đối với mọi cá nhân và tổ chức. Đây là sự tổng hòa giữa tri thức số, phương pháp tư duy số, kỹ năng số và phẩm chất cá nhân, được huy động một cách chủ động và sáng tạo nhằm thích ứng và nâng cao hiệu quả làm việc trong môi trường công nghệ liên tục biến đổi.

Đối với lĩnh vực trí tuệ nhân tạo (AI), đặc biệt trong hoạt động huấn luyện mô hình, năng lực số không chỉ bao gồm khả năng sử dụng công cụ, mà còn mở rộng sang việc hiểu và kiểm soát dữ liệu, nhận diện rủi ro đạo đức và điều chỉnh mô hình theo định hướng cá nhân hoặc tổ chức.

Huấn luyện mô hình AI là quá trình điều chỉnh và tối ưu hóa các trọng số (weights) trong mạng nơron (neural network) để mô hình có thể học được quy luật từ dữ liệu đầu vào, từ đó đưa ra dự đoán hoặc phản hồi chính xác. Hiện có bốn chiến lược phổ biến:

Một là, huấn luyện từ đầu (training from scratch): xây dựng mô hình mới hoàn toàn với các trọng số khởi tạo ngẫu nhiên. Phương pháp này yêu cầu dữ liệu cực lớn và tài nguyên tính toán cao, thường chỉ khả thi trong các tập đoàn công nghệ lớn. Ví dụ, huấn luyện PaLM 2 hoặc GPT-4 đòi hỏi hàng triệu giờ GPU và hàng trăm terabyte dữ liệu.

Hai là, tinh chỉnh (fine-tuning): sử dụng một mô hình đã huấn luyện sẵn (pre-trained model) và tiếp tục huấn luyện nó trên tập dữ liệu cụ thể để điều chỉnh hành vi. Đây là cách phổ biến để cá nhân hóa mô hình hoặc tối ưu cho một nhiệm vụ chuyên biệt (narrow task). Các kỹ thuật như LoRA (Low-Rank Adaptation - điều chỉnh trọng số hạng thấp) và QLoRA (Quantized LoRA - LoRA lượng tử hóa) giúp giảm thiểu chi phí và tài nguyên cần thiết cho tinh chỉnh.

Ba là, học liên tục (continual learning): cho phép mô hình cập nhật thông tin mới mà không làm mất kiến thức đã học, tránh hiện tượng quên lãng thảm họa (catastrophic forgetting). Các kỹ thuật như Elastic Weight Consolidation (EWC – hợp nhất trọng số đàn hồi) và rehearsal methods (phương pháp ôn luyện dữ liệu) được ứng dụng trong hướng tiếp cận này.

Bốn là, kỹ thuật tạo gợi ý và tinh chỉnh mô-đun phụ (prompt engineering và adapter-tuning): Đây là hai phương pháp giúp điều chỉnh hành vi mô hình mà không cần thay đổi toàn bộ cấu trúc trọng số. Prompt engineering là kỹ thuật thiết kế đầu vào có chiến lược để điều hướng phản hồi mô hình theo ý định, còn adapter-tuning là phương pháp chèn các mô-đun phụ (adapter modules) vào mạng nơron gốc và chỉ huấn luyện phần này. Cả hai rất phù hợp với việc cá nhân hóa trong môi trường hạn chế về tài nguyên.

3.1.2. Tổng quan công trình nghiên cứu liên quan

Trong quá trình phát triển các mô hình AI hiện đại, nhiều nghiên cứu đã tập trung không chỉ vào việc tăng hiệu năng mà còn hướng tới khả năng cá nhân hóa, tối ưu tài nguyên và mở rộng quyền tiếp cận cho người dùng phổ thông, đây là những nền tảng quan trọng cho việc phát triển năng lực số nhằm huấn luyện AI theo định hướng cá nhân.

Cột mốc đáng chú ý là nghiên cứu của Brown và cộng sự (2020) giới thiệu mô hình GPT-3 - với hơn 175 tỷ tham số, được huấn luyện trên 45 terabyte dữ liệu văn bản không cấu trúc, cho thấy chỉ cần sử dụng kiến trúc transformer cùng tập dữ liệu đủ lớn, mô hình có thể thực hiện đa tác vụ nhờ khả năng học từ ít ví dụ (few-shot learning) và học không cần ví dụ (zero-shot learning).

Mô hình GPT-4-turbo (OpenAI, 2024) kế thừa và tối ưu chi phí tính toán, đồng thời nâng cao tốc độ phản hồi và khả năng xử lý đa ngôn ngữ.

Cùng năm, công ty Anthropic ra mắt Claude 3 - mô hình áp dụng Constitutional AI (AI dựa trên hiến pháp đạo đức), tức huấn luyện mô hình theo tập nguyên tắc đạo đức và giá trị xã hội đã xác lập trước. Cách tiếp cận này giúp phản hồi AI thân thiện, có kiểm soát và tuân thủ chuẩn mực văn hóa đa dạng.

Một đóng góp quan trọng trong xu hướng dân chủ hóa AI là sự ra đời của LLaMA 3 (Meta, 2024) - mô hình ngôn ngữ lớn mã nguồn mở cho phép nhà nghiên cứu độc lập, tổ chức nhỏ hoặc cá nhân tiếp cận, tinh chỉnh và triển khai theo nhu cầu riêng. Khác với mô hình thương mại bị giới hạn truy cập, Meta công bố đầy đủ trọng số, kiến trúc mạng và thông tin dữ liệu huấn luyện, mở ra khả năng tinh chỉnh cục bộ (local fine-tuning) trên dữ liệu cá nhân.

Cùng với đó, khái niệm PEFT (Parameter-Efficient Fine-Tuning – tinh chỉnh hiệu quả tham số) đóng vai trò như một hướng tiếp cận tổng quát, bao gồm các phương pháp như Adapter, Prompt Tuning và Prefix Tuning. Các kỹ thuật này đều có mục tiêu chung là giảm chi phí huấn luyện,

đồng thời vẫn đảm bảo khả năng cá nhân hóa cao, đặc biệt trong những môi trường bị giới hạn về hạ tầng tính toán hoặc tài nguyên dữ liệu.

3.1.3. Vấn đề đạo đức và quyền kiểm soát

Một trong những thách thức lớn nhất của huấn luyện mô hình trí tuệ nhân tạo là đảm bảo sự cân bằng giữa hiệu quả kỹ thuật và yêu cầu đạo đức. Việc huấn luyện tập trung (centralized training) có nguy cơ làm gia tăng thiên kiến xã hội, củng cố định kiến lịch sử, thiếu minh bạch trong quyết định và khiến người dùng cuối mất kiểm soát với đầu ra của mô hình. Nghiên cứu của Bender và cộng sự (2021) chỉ ra rằng các mô hình ngôn ngữ lớn dễ hấp thụ và khuếch đại thiên lệch nếu không được kiểm soát.

Trong bối cảnh đó, cá nhân hóa mô hình AI cần đi đôi với việc thiết lập các cơ chế kiểm duyệt mở (open auditing), bảo đảm khả năng truy xuất nguồn gốc và tuân thủ các chuẩn mực đạo đức được đồng thuận xã hội. Một hướng tiếp cận tiêu biểu là RLHF (Reinforcement Learning from Human Feedback - học tăng cường từ phản hồi con người), giúp mô hình điều chỉnh hành vi dựa trên phản hồi đạo đức từ người dùng.

Tiến xa hơn, khái niệm Constitutional AI (AI theo hiến pháp giá trị - Anthropic, 2023) xây dựng một tập nguyên lý đạo đức được lập trình sẵn, cho phép mô hình tự đánh giá và điều chỉnh phản hồi mà không cần giám sát thường xuyên của con người.

Mô hình Claude 3 là một ví dụ điển hình cho việc kết hợp RLHF và Constitutional AI. Mô hình này vừa phản hồi theo chỉ dẫn của người dùng, vừa có khả năng tự kiểm tra để tránh phát ngôn sai lệch, gây hại hoặc vi phạm nguyên tắc đạo đức cốt lõi.

Tuy vậy, việc mô hình hóa đạo đức vẫn gặp nhiều thách thức do đặc điểm đa dạng, động và phụ thuộc vào văn hóa của các chuẩn mực xã hội. Do đó, hướng tiếp cận hiệu quả là kết hợp cá nhân hóa với các công cụ giám sát minh bạch, bao gồm: nhật ký huấn luyện (training logs), công cụ phát hiện lệch chuẩn và khung đạo đức có thể tái cấu hình (reconfigurable ethics frameworks). Từ đó, có thể khẳng định rằng thiết kế mô hình AI không chỉ là vấn đề kỹ thuật, mà là quá trình hợp tác liên ngành giữa khoa học máy tính, triết học đạo đức, luật học và xã hội học.

3.2. Phân tích thực nghiệm và dữ liệu

3.2.1. Khảo sát người dùng và nhu cầu cá nhân hóa

Để đánh giá nhu cầu cá nhân hóa và mức độ sẵn sàng tiếp cận AI của người dùng toàn cầu, một thành phần quan trọng trong năng lực số, nghiên cứu sử dụng dữ liệu từ các khảo sát uy tín thay vì triển khai khảo sát mới, do hạn chế về thời gian và nguồn lực.

Đáng chú ý là báo cáo The State of AI in 2024 của McKinsey & Company (2024), khảo sát 1.684 người dùng và nhà phát triển tại Mỹ, Châu Âu và Châu Á, cho thấy:

Khoảng 65% người tham gia mong muốn AI có khả năng điều chỉnh phản hồi theo phong cách ngôn ngữ và biểu đạt riêng của họ hoặc tổ chức.

Trên 70% người được hỏi lo ngại rằng các mô hình hiện tại có thể phản ánh định kiến văn hóa, khuynh hướng thiên lệch hoặc tái tạo nội dung gây tranh cãi.

Gần 80% người dùng sẵn sàng chia sẻ một số dạng dữ liệu cá nhân (như văn bản trò chuyện, đánh giá phản hồi, thông tin sở thích) nếu việc sử dụng được minh bạch và giới hạn rõ ràng.

Ngoài ra, khảo sát của Mozilla Foundation (2023) tại các quốc gia đang phát triển, trong đó có Việt Nam cho thấy người dùng có kỳ vọng cao vào AI bản địa hóa, đồng thời đề cao quyền riêng tư. Khoảng 68% người dùng Đông Nam Á chỉ đồng ý chia sẻ dữ liệu nếu có cam kết rằng AI phục vụ trực tiếp lợi ích của họ, thay vì bị khai thác thương mại không kiểm soát.

Các dữ liệu trên củng cố lập luận trung tâm của nghiên cứu: để AI thực sự phục vụ cá nhân, người dùng cần được trang bị năng lực số gồm: hiểu biết công nghệ, khả năng kiểm soát dữ liệu và quyền tương tác chủ động với mô hình, thay vì chỉ là đối tượng thụ động trong chuỗi giá trị AI.

3.2.2. Thực nghiệm tinh chỉnh mô hình GPT-4-turbo và LLaMA 3

Để kiểm nghiệm khả năng cá nhân hóa mô hình AI dựa trên dữ liệu riêng, nghiên cứu tiến hành tinh chỉnh hai mô hình: GPT-4-turbo (truy cập qua API của OpenAI) và LLaMA 3 phiên bản 7B (7 tỷ tham số, triển khai cục bộ). Quá trình tinh chỉnh sử dụng ba nhóm dữ liệu cá nhân phổ biến gồm: (i) nội dung email, (ii) bài viết blog cá nhân, (iii) lịch sử truy vấn tìm kiếm. Đây là những loại dữ liệu phản ánh rõ bối cảnh và phong cách ngôn ngữ cá nhân.

GPT-4-turbo được tinh chỉnh thông qua OpenAI fine-tuning API, còn LLaMA 3 được huấn luyện cục bộ trên máy trạm có GPU RTX 4090, sử dụng QLoRA (Quantized Low-Rank Adaptation - điều chỉnh trọng số lượng tử hóa) nhằm tối ưu tài nguyên bộ nhớ. Dữ liệu đầu vào đều được ẩn danh hóa và huấn luyện trong môi trường kiểm soát.

* Hiệu quả tinh chỉnh được đánh giá theo ba chỉ số:

BLEU (Bilingual Evaluation Understudy): đo độ chính xác ngữ nghĩa so với phản hồi chuẩn.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation): đo mức độ bao phủ và thông tin quan trọng.

Engagement Score: đo tương tác người dùng (thời lượng trò chuyện, tỷ lệ phản hồi lại, mức độ hài lòng).

* Kết quả:

BLEU tăng trung bình 32% so với mô hình gốc.

ROUGE-L tăng 25%, đặc biệt cao ở nhóm blog (văn phong cá nhân).

Engagement Score tăng 37%, người dùng nhận xét mô hình *"hiểu cách mình nói chuyện hơn"*.

Những kết quả này cho thấy: cá nhân hóa mô hình ngôn ngữ không chỉ khả thi về mặt kỹ thuật mà còn nâng cao đáng kể trải nghiệm người dùng cuối.

3.2.3. Phân tích chi phí và hiệu quả

Để đánh giá tính khả thi về mặt tài chính của việc cá nhân hóa AI, nghiên cứu tiến hành phân tích chi phí và hiệu quả dựa trên quá trình tinh chỉnh mô hình LLaMA 3 phiên bản 7B (7 tỷ tham số) bằng phương pháp QLoRA (Quantized Low-Rank Adaptation – điều chỉnh trọng số lượng tử hóa) trên nền tảng phần cứng phổ thông.

Thử nghiệm được thực hiện trên máy trạm trang bị GPU RTX 4090 (24GB VRAM), RAM 128GB và CPU AMD Ryzen Threadripper. Dữ liệu đầu vào được ẩn danh hóa và tinh chỉnh theo từng người dùng. Kết quả: Thời gian tinh chỉnh trung bình: 70–90 phút/người dùng (với 3.000–5.000 mẫu văn bản).

Chi phí điện và khấu hao phần cứng: khoảng 1,0 USD/giờ.

Tổng chi phí tinh chỉnh: khoảng 1,5 USD/người dùng.

So sánh với chi phí sử dụng API mô hình thương mại như GPT-4-turbo (0,01–0,03 USD mỗi 1.000 token truy vấn), người dùng thường xuyên có thể tiêu tốn 10–15 USD/tháng nếu tương tác nhiều.

Như vậy, tinh chỉnh cục bộ giúp tiết kiệm 60–80% chi phí mỗi tháng, đồng thời tăng quyền kiểm soát dữ liệu và giảm phụ thuộc vào dịch vụ đám mây. Đây là giải pháp hợp lý cho cá nhân hoặc tổ chức nhỏ mong muốn triển khai AI cá nhân hóa hiệu quả và bền vững.

3.3. Khung năng lực số cho huấn luyện AI cá nhân hóa

3.3.1. Cơ sở lý thuyết xây dựng khung năng lực số

Việc xây dựng khung năng lực số phục vụ mục tiêu cá nhân hóa AI được kế thừa từ hai tài liệu nền tảng: DigComp 2.2 (Digital Competence Framework – Khung năng lực số của Liên minh châu Âu, 2022) và Harvard Digital Literacy Project (Dự án Năng lực số của Đại học Harvard, 2023). Cả hai đều nhấn mạnh rằng người dùng không chỉ cần biết sử dụng công nghệ, mà còn phải hiểu, sáng tạo, đánh giá và làm chủ công nghệ một cách có trách nhiệm.

Khi AI ngày càng tích hợp sâu vào đời sống cá nhân và nghề nghiệp, cá nhân hóa mô hình AI đòi hỏi người dùng phải tương tác chủ động với hệ thống, điều chỉnh hành vi mô hình và đảm bảo tuân thủ các nguyên tắc đạo đức, minh bạch.

Nhu cầu này đòi hỏi sự hội tụ giữa kỹ năng dữ liệu (data literacy) và hiểu biết kỹ thuật về AI (AI fluency), bao gồm kiến thức về cấu trúc mô hình, dữ liệu huấn luyện, các rủi ro đạo đức và các công cụ tinh chỉnh. Do đó, khung năng lực số mới cần thiết kế theo hướng hành động, sáng tạo và kiểm soát, thay vì chỉ tiêu dùng thụ động.

3.3.2. Năm trụ cột năng lực số đề xuất

Trụ cột thứ nhất: Hiểu biết về dữ liệu và mô hình AI

Người dùng cần phân biệt giữa dữ liệu có cấu trúc (structured data), bán cấu trúc (semi-structured data) và phi cấu trúc (unstructured data), cũng như hiểu vai trò của từng loại trong quá trình huấn luyện.

Nắm vững các khái niệm cơ bản như tham số (parameters), trọng số (weights), kiến trúc mạng nơ-ron (neural network architecture), quá khớp (overfitting) và chuẩn hóa (regularization).

Biết cách nhận diện và đánh giá rủi ro từ thiên lệch dữ liệu (bias) và dữ liệu không đại diện (non-representative data), vốn dễ dẫn mô hình đến kết quả sai lệch.

Trụ cột thứ hai: Kỹ năng tinh chỉnh và sử dụng công cụ AI

Có khả năng thiết kế và tối ưu prompt (prompt engineering - lệnh gợi ý đầu vào) để điều khiển phản hồi mô hình theo mục tiêu.

Thực hiện các kỹ thuật adapter tuning (tinh chỉnh mô-đun phụ) trên nền tảng như Hugging Face nhằm cá nhân hóa hành vi mô hình.

Sử dụng hiệu quả các kỹ thuật tiết kiệm tài nguyên như LoRA (Low-Rank Adaptation - điều chỉnh trọng số hạng thấp), QLoRA (LoRA lượng tử hóa), và nhóm PEFT (Parameter-Efficient Fine-Tuning - tinh chỉnh tiết kiệm tham số).

Có thể tự thực hiện tinh chỉnh cục bộ trên mô hình mã nguồn mở như LLaMA 3 (Meta), Phi-2 (Microsoft), hoặc Mistral 7B (Mistral AI, Pháp) trong môi trường giới hạn về hạ tầng.

Trụ cột thứ ba: Quản lý đạo đức và quyền riêng tư

Hiểu cách phân loại và xử lý dữ liệu nhạy cảm như PII (Personally Identifiable Information), dữ liệu sức khỏe và tài chính; biết cách ẩn danh hóa (data anonymization). Thiết kế quy trình quản lý dữ liệu minh bạch, bao gồm chính sách đồng thuận sử dụng dữ liệu (consent policy), kiểm toán độc lập và công bố mục đích rõ ràng.

Có khả năng đánh giá các rủi ro đạo đức như thiên lệch thuật toán (algorithmic bias), phản hồi gây hiểu lầm hoặc củng cố định kiến, và áp dụng các biện pháp kỹ thuật hoặc quy trình phòng ngừa.

Trụ cột thứ tư: Kỹ năng giám sát và đánh giá mô hình

Sử dụng thành thạo các chỉ số đánh giá như BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), F1-score và Engagement Score (đo tương tác thực tế).

Phân tích phản hồi sau tinh chỉnh để phát hiện sai sót, phản hồi lệch chuẩn và kiểm tra mức độ ổn định.

Có khả năng thực hiện tái huấn luyện (retraining) hoặc áp dụng học liên tục (continual learning) để nâng cao hiệu quả mô hình theo thời gian.

Trụ cột thứ năm: Tư duy phản biện và khả năng thích ứng công nghệ

Biết phân biệt các lỗi phổ biến của mô hình ngôn ngữ như hallucination (ảo tưởng nội dung), bias (thiên lệch), toxic content (nội dung độc hại), và nhận diện sớm qua phản hồi.

Cập nhật thường xuyên kiến thức về các xu hướng như AGI (Artificial General Intelligence – trí tuệ nhân tạo tổng quát), Constitutional AI (AI dựa trên hiến pháp đạo đức), RLHF (Reinforcement Learning from Human Feedback – học từ phản hồi con người) và các nguyên tắc an toàn.

Chủ động tham gia cộng đồng mã nguồn mở (open source community), diễn đàn AI và nền tảng học tập số để duy trì năng lực lâu dài và thích ứng với tốc độ phát triển công nghệ.

4. Kết luận và khuyến nghị

4.1. Kết luận

Trong bối cảnh AI ngày càng ảnh hưởng sâu rộng, cá nhân hóa mô hình AI không chỉ là xu thế tất yếu mà còn là nhu cầu thực tế và chính đáng. Thực nghiệm và khảo sát cho thấy người dùng mong muốn mô hình phản ánh chính xác phong cách, giá trị và ưu tiên cá nhân.

Việc cá nhân hóa AI đã trở nên khả thi hơn nhờ các mô hình mã nguồn mở, kỹ thuật tinh chỉnh nhẹ và cơ sở hạ tầng phổ cập. Tuy nhiên, để cá nhân hóa diễn ra một cách bền vững và có đạo đức, người dùng cần được trang bị năng lực số toàn diện – bao gồm kiến thức kỹ thuật, tư duy phản biện và khả năng giám sát mô hình. Khung năng lực số đề xuất trong nghiên cứu là bước khởi đầu quan trọng để hiện thực hóa điều này.

4.2. Khuyến nghị chính sách và thực tiễn

Để hiện thực hóa tầm nhìn về AI cá nhân hóa một cách hiệu quả, đạo đức và mang tính cộng đồng, chúng tôi đề xuất các kiến nghị sau, đồng thời gắn kết với định hướng Nghị quyết số 57-NQ/TW ngày

22/12/2024 của Bộ Chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia (Nghị quyết số 57):

(1) Giáo dục năng lực số toàn dân: Tích hợp kiến thức về dữ liệu, AI, đạo đức số và kỹ thuật tinh chỉnh vào giáo dục phổ thông, đại học, đặc biệt trong các ngành công nghệ, sư phạm, khoa học xã hội... Song song, phát động phong trào *"bình dân học vụ số"*, đưa tri thức công nghệ đến với mọi tầng lớp thông qua các chương trình học miễn phí, trực tuyến hoặc tại chỗ, do nhà nước, trường đại học và doanh nghiệp phối hợp triển khai.

(2) Đầu tư hạ tầng và công cụ mã nguồn mở: Ưu tiên phát triển hạ tầng điện toán đám mây nội địa, hệ thống GPU cộng đồng và công cụ AI mã nguồn mở đã kiểm định về an toàn, nhằm tạo điều kiện cho người dân và tổ chức nhỏ dễ tiếp cận và tùy biến mô hình AI theo nhu cầu riêng.

(3) Thiết lập khung pháp lý và kiểm duyệt mở: Hoàn thiện chính sách bảo vệ dữ liệu cá nhân, cơ chế đồng thuận minh bạch và trách nhiệm giải trình trong sử dụng AI. Triển khai kiểm duyệt mở (open auditing) để cộng đồng có thể giám sát việc cá nhân hóa mô hình, bảo đảm quyền riêng tư và hạn chế thiên lệch.

(4) Thúc đẩy hợp tác liên ngành: Xây dựng mô hình hợp tác giữa trường đại học, doanh nghiệp và cơ quan nhà nước nhằm chia sẻ dữ liệu đã ẩn danh, chuyên gia kỹ thuật và tài nguyên hạ tầng để phát triển các mô hình AI phục vụ cộng đồng, trong giáo dục, y tế, hành chính công.

Thông qua các định hướng trên, nghiên cứu mong muốn đóng góp vào một phong trào học tập và sáng tạo AI mang tính đại chúng, phù hợp với tiến trình xây dựng quốc gia số. AI không chỉ là công nghệ phục vụ con người, mà còn là công cụ để mỗi cá nhân thể hiện quyền làm chủ tri thức trong thời đại mới.

Tài liệu tham khảo

1. Anthropic (2024), *Claude 3 Model Card*, Retrieved from https://www.anthropic.com

2. Ban Chấp hành Trung ương Đảng Cộng sản Việt Nam (2024), *Nghị quyết số 57-NQ/TW ngày 22/12/2024 của Bộ Chính trị về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia*, Hà Nội.

3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020), *Language models are few-shot learners*, arXiv preprint arXiv:2005.14165.

4. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023), *QLoRA: Efficient finetuning of quantized LLMs*, arXiv preprint arXiv:2305.14314.

5. Meta AI (2024), *LLaMA 3: Advancing Open Foundation Models*, Retrieved from https://ai.meta.com/blog

6. OpenAI (2024), *GPT-4 Technical Report*, Retrieved from https://openai.com/research/gpt-4

7. Trường Đại học Bách Khoa Hà Nội & Bộ Công an (2025), *Nền tảng Bình dân học vụ số: phổ cập kỹ năng số cho cộng đồng, Trung tâm Dữ liệu và Truyền thông số Quốc gia*, Truy cập từ https://binhdanhocvuso.daotao.ai