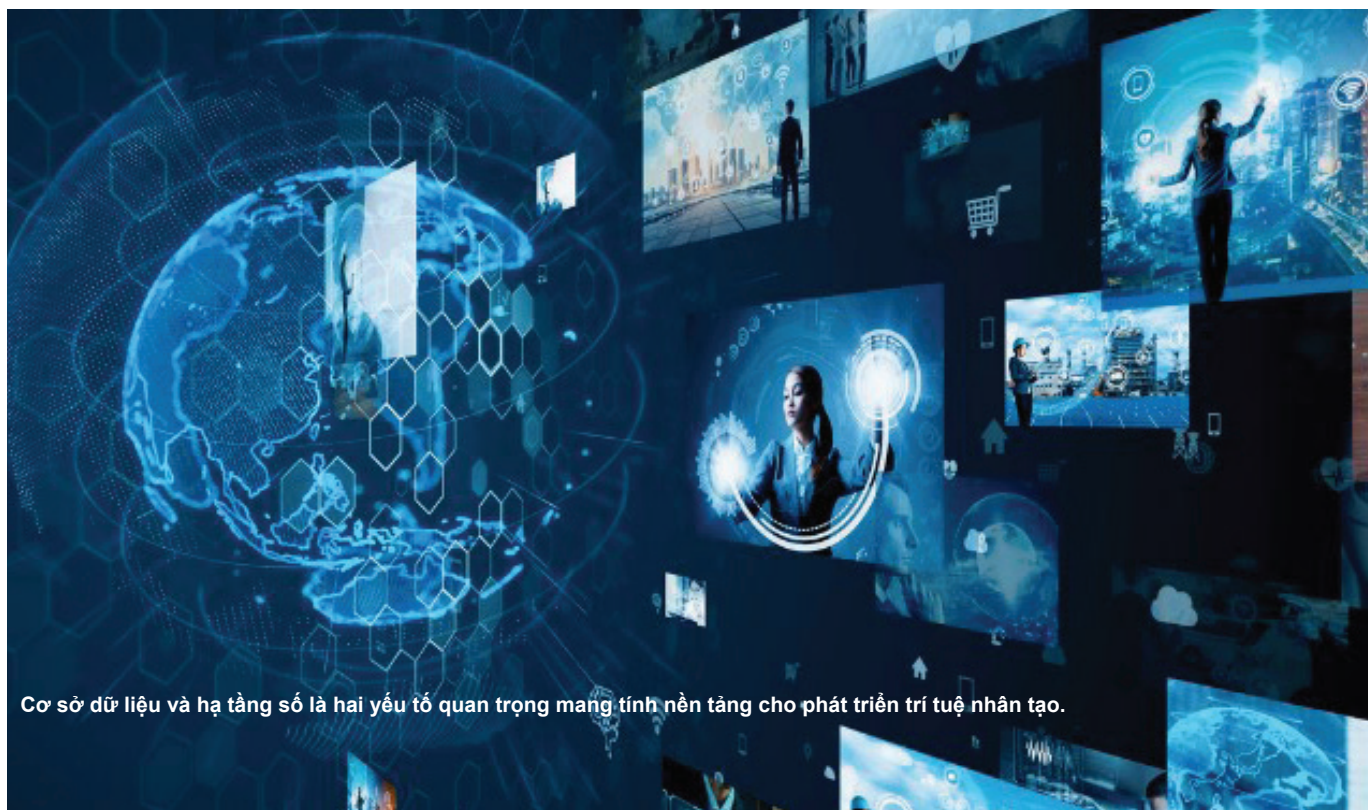




Cơ sở dữ liệu và hạ tầng số là hai yếu tố quan trọng mang tính nền tảng cho phát triển trí tuệ nhân tạo.

DỮ LIỆU VÀ HẠ TẦNG SỐ: Nền tảng vững chắc cho phát triển trí tuệ nhân tạo

Nguyễn Lan Phương¹, Vũ Văn Hưng²

¹Viện Nghiên cứu Chính sách và Phát triển Truyền thông (IPS)

²Tạp chí Khoa học và Công nghệ Việt Nam

“

Cơ sở dữ liệu (CSDL) và hạ tầng số là hai yếu tố quan trọng, mang tính nền tảng cho phát triển trí tuệ nhân tạo (AI). Trong các cơ quan Nhà nước, CSDL khá phong phú, hạ tầng số bước đầu được xây dựng và phát triển. Tuy nhiên, CSDL và hạ tầng số trong các cơ quan Nhà nước đang gặp những khó khăn và thách thức, đòi hỏi cần có những giải pháp khả thi cho trước mắt và cả lâu dài.

”

Cơ sở dữ liệu được xây dựng và phát triển mạnh mẽ

Các cơ quan Nhà nước hiện nay đang sở hữu một lượng lớn dữ liệu chất lượng cao, tính chất chuyên môn sâu, phạm vi bao phủ rộng, tạo cơ sở vững chắc cho các chức năng khác nhau của chính phủ và quy trình hoạch định chính sách. Việc thu thập toàn diện dữ liệu trong các cơ quan Nhà nước mang lại cơ hội đáng kể cho các ứng dụng nâng cao, đặc biệt là AI. Hiểu được cơ hội này, các cơ quan Nhà nước sẽ xác định được tầm quan trọng của dữ liệu chất lượng cao đối với việc triển khai thành công các ứng dụng AI.

Năm 2024, tổng số CSDL quốc gia ưu tiên triển khai là 10 (trong đó 05 CSDL quốc gia đã hoàn thành, khai thác sử dụng; 03 CSDL đang triển khai; 02 CSDL đã có kế hoạch, bắt đầu triển khai). Dữ liệu dân cư quốc gia đạt mức độ hoàn chỉnh, tạo điều kiện kết nối, chia sẻ, sử dụng dữ liệu cả trong và ngoài khu vực nhà nước. Về dữ liệu mở của khu vực Nhà nước, năm 2024 tổng số có 75 cơ quan, địa phương đã ban hành danh mục, kế hoạch cung cấp dữ liệu mở (chiếm 87%); hơn 30 cổng dữ liệu mở đã được xây dựng (bao gồm cả thử nghiệm). Ngoài các CSDL quốc gia, các CSDL chuyên ngành khác nhau, CSDL của địa phương cũng được phát triển. Những nỗ lực đáng kể đã được thực hiện để tạo điều kiện thuận lợi cho việc xử lý và trao đổi dữ liệu thông suốt trong và giữa các cơ quan, cũng như giữa các cấp chính quyền khác nhau, từ trung ương đến địa phương.

Ở khối doanh nghiệp, các “ông lớn” như Tập đoàn Công nghiệp Viễn thông Quân đội (Viettel), Tập đoàn Bưu chính Viễn thông Việt Nam (VNPT), Công ty Cổ phần FPT, các ngân hàng thương mại... từ lâu đã chuẩn hóa dữ liệu của họ để tận dụng các nguồn dữ liệu nhằm tối ưu hóa hoạt động, chức năng và dịch vụ khách hàng của họ. Các tổ chức này đã đầu tư đáng kể vào số hóa dữ liệu và đã thiết lập các phương pháp quản lý dữ liệu mạnh mẽ để hỗ trợ các mục tiêu chiến lược và nâng cao hiệu quả hoạt động.

Dữ liệu, hạ tầng liên quan đến ứng dụng trí tuệ nhân tạo

Thách thức, khó khăn về dữ liệu

Độ trưởng thành và mức độ sẵn sàng của dữ liệu khác nhau đáng kể giữa các lĩnh vực và các cơ quan nhà nước đang là thách thức lớn. Các cuộc phỏng vấn được thực hiện ở các cơ quan trung ương, cũng như các địa phương cho thấy, dữ liệu còn rời rạc, phân mảnh; việc chia sẻ, kết nối dữ liệu giữa các cơ quan, các cấp còn hạn chế khiến các cơ quan không thể sử dụng đầy đủ nguồn dữ liệu phong phú có sẵn này; nhất là dữ liệu ngành dọc đang khép kín, chưa được chia sẻ nhiều. Việc thiếu dữ liệu sạch, đã được xử lý cũng là một rào cản lớn. Vẫn tồn tại một lượng lớn dữ liệu không đầy đủ hoặc cần được xử lý đáng kể trước khi được sử dụng vào phát triển, huấn luyện mô hình, giải pháp AI. Ngoài ra, nhiều bộ dữ liệu cũ của các cơ quan nhà nước vẫn chưa được số hóa và phần lớn được lưu trữ cục bộ, rời rạc, thậm chí ở các định dạng

vật lý, chẳng hạn như tài liệu giấy. Điều này càng làm phức tạp thêm đối với việc sử dụng dữ liệu cho các ứng dụng AI.



Các doanh nghiệp vừa và nhỏ gặp nhiều thách thức trong quản lý dữ liệu.

Ở khu vực tư, các doanh nghiệp vừa và nhỏ (SMEs) gặp nhiều thách thức trong quản lý dữ liệu. SMEs thiếu năng lực kỹ thuật để số hóa và tối ưu hóa sử dụng dữ liệu hiện có, khó tiếp cận các nguồn dữ liệu về thị trường, dữ liệu từ các cơ quan Nhà nước, cũng như không chắc chắn về chi phí - lợi ích của việc áp dụng các giải pháp kỹ thuật số ở giai đoạn này. Do đó, nhiều SMEs vẫn chưa hoàn toàn sẵn sàng về dữ liệu để ứng dụng AI nội bộ, chứ chưa nói đến phát triển các giải pháp AI cho các bên, trong đó có các cơ quan Nhà nước.

Thiếu dữ liệu mở đủ lớn, đáp ứng chất lượng, nhất là dữ liệu mở từ các cơ quan Nhà nước là một điểm bất cập, làm hạn chế khả năng tiếp cận dữ liệu của doanh nghiệp, các viện nghiên cứu, trường đại học và bản thân các cơ quan nhà nước để phát triển các giải pháp, ứng dụng AI, vốn cần các nguồn dữ liệu đa dạng, có chất lượng.

Thách thức, khó khăn về hạ tầng

Khi nhu cầu đối với các ứng dụng AI mạnh ngày càng nhiều trong các cơ quan Nhà nước, cần có các thiết bị cấu hình cao và tiêu tốn công suất gấp 3-4 lần so với hạ tầng như hiện tại. Các mô hình, ứng dụng AI như vậy đòi hỏi lượng tài nguyên tính toán, xử lý dữ liệu rất lớn, hạ tầng máy chủ phải trang bị từ GPU tới NPU hiệu suất cao, cho phép xử lý hàng nghìn tỷ phép tính mỗi giây. Hơn nữa, dòng đời hạ tầng phần cứng ngày càng ngắn (chỉ kéo dài 1-2 năm so với



Trung tâm dữ liệu Viettel Hoà Lạc (Nguồn: VGP/MT).

3-5 năm trước đây), ví dụ như các dòng card GPU được NVIDIA H100 phát hành tháng 3/2023, đến giữa năm 2024 đã có thêm 3 version cao cấp hơn: H200, Blackwell B100, Blackwell B200 với hiệu suất cao, giá cạnh tranh.

Một rào cản đáng kể để thúc đẩy việc áp dụng AI trong các cơ quan nhà nước là thiếu cơ sở hạ tầng tiên tiến, cập nhật cho các ứng dụng AI phức tạp hơn. Tại thời điểm cuối năm 2024, tổng công suất của các Trung tâm dữ liệu ở Việt Nam khoảng hơn 50 MW, dự kiến thêm 11 MW từ các trung tâm dữ liệu đang được xây dựng, còn thấp so với các nước trong khu vực. Bên cạnh đó, do thiếu nguồn lực và cơ chế tài chính, cũng như bối cảnh cạnh tranh công nghệ trên thế giới, rất khó khăn để có chip AI (ví dụ: H100, chưa nói đến các chip AI đời sau) và trung tâm dữ liệu AI chuyên dụng, rất cần thiết để xử lý các bộ dữ liệu lớn và chạy các mô hình AI sử dụng nhiều tài nguyên.

Một giải pháp được một số cơ quan nhà nước áp dụng là phối hợp tận dụng nguồn lực hạ tầng của các doanh nghiệp lớn như Viettel, FPT. Tuy nhiên, các doanh nghiệp bên ngoài thường cung cấp các bộ cơ sở hạ tầng được tiêu chuẩn hóa, có sẵn, có thể chưa tính đến đầy đủ đặc thù, nhu cầu của cơ quan Nhà nước, các quy định, cũng như các yêu cầu chuyên môn cụ thể ở ngành, lĩnh vực cần ứng dụng AI. Sự không phù hợp giữa các giải pháp được cung cấp và các yêu cầu thực tế của các cơ quan có thể dẫn đến hiệu quả không cao, việc triển khai AI không tối ưu, không đáp ứng các mục tiêu đặt ra.

Một vấn đề quan trọng có thể nảy sinh trong việc đảm bảo cơ sở hạ tầng lõi cho phát triển AI trong các cơ quan nhà nước và tư nhân là tình trạng thiếu năng lượng vận hành các cơ sở này, đặc biệt là các trung tâm dữ liệu chuyên dụng cho AI. Trong giai đoạn từ năm 2010 đến 2022, công suất tiêu thụ điện của các trung tâm dữ liệu tại Việt Nam đã tăng gấp ba lần, trong khi băng thông sử dụng tăng từ 10 đến 15 lần. Chẳng hạn, trung tâm dữ liệu lớn nhất Việt Nam - Trung tâm dữ liệu thứ 14 của Tập đoàn Viettel đặt tại Khu Công nghệ cao Hòa Lạc có công suất điện tiêu thụ lên tới 30 MW. Tổng cộng 14 trung tâm dữ liệu của Viettel có

công suất tổng lên đến 87 MW, tương đương với mức tiêu thụ điện của gần 70.000 hộ gia đình Bắc Mỹ. Dự kiến các doanh nghiệp sẽ đầu tư xây dựng thêm nhiều trung tâm dữ liệu. Với hệ thống điện hiện tại khó đáp ứng được nhu cầu tiêu thụ điện cao điểm từ tháng 5 đến tháng 7, các trung tâm dữ liệu vận hành liên tục có thể trở thành một gánh nặng đáng kể cho lưới điện quốc gia.

Khuyến nghị về dữ liệu, hạ tầng cho ứng dụng trí tuệ nhân tạo

Tối đa hóa lợi ích của dữ liệu, hạ tầng hiện có cho ứng dụng AI: Những trường hợp đã ứng dụng AI trong các cơ quan Nhà nước cho thấy một số bài học thiết thực: Dữ liệu của các cơ quan Nhà nước đáp ứng yêu cầu đối với những giải pháp AI nhất định; có thể tận dụng lợi thế này để mở rộng việc xây dựng các giải pháp AI. Cách tiếp cận thực tế, tùy theo điều kiện về dữ liệu để xây dựng các ứng dụng AI là phù hợp trong giai đoạn hiện nay, ví dụ như trợ lý ảo cho cán bộ, công chức; công cụ AI giúp soạn thảo văn bản đơn giản; giải pháp AI hỗ trợ rà soát văn bản; giải pháp AI rà soát, đơn giản hóa các bước của thủ tục hành chính.

Phát triển nguồn dữ liệu: Để khai thác tối đa tiềm năng của dữ liệu cho việc phát triển, ứng dụng AI trong các cơ quan Nhà nước, cần có nhiều nỗ lực hơn nữa để tích hợp và số hóa dữ liệu giấy, đảm bảo rằng dữ

liệu sau khi số hóa sẽ được tiêu chuẩn hóa, chia sẻ, sẵn sàng sử dụng trong các ứng dụng AI. Giữa các cơ quan nhà nước với nhau, cần tiếp tục xây dựng, mở rộng các CSDL dùng chung. Có thể là chuyển đổi CSDL hiện có của một ngành thành CSDL dùng chung, cấp quyền truy cập theo chức năng, nhiệm vụ của các cơ quan...

Kết nối, chia sẻ dữ liệu: Để kết nối, chia sẻ dữ liệu, cần xây dựng các tiêu chuẩn/quy chuẩn/quy định về cấu trúc dữ liệu trao đổi đối với dữ liệu trong CSDL dùng chung của bộ, ngành, địa phương; chuẩn hóa các dịch vụ chia sẻ dữ liệu (API) từ CSDL dùng chung và cung cấp trên nền tảng dịch vụ chính phủ (LGSP); kết nối các hệ thống thông tin, ứng dụng của các đơn vị với các dịch vụ chia sẻ dữ liệu trên LGSP. Cần xác định các dữ liệu có thể được chia sẻ ra bên ngoài bộ, ngành, địa phương (chia sẻ cho các bộ, ngành khác) để cung cấp trên nền tảng tích hợp. Các trung tâm dữ liệu cần chia sẻ dữ liệu cho cả các cơ quan Nhà nước, doanh nghiệp AI và khối nghiên cứu trong việc phát triển các giải pháp AI cho các cơ quan Nhà nước.

Đơn giản hóa thủ tục hành chính: Các thủ tục hành chính liên quan đến dữ liệu cần được đơn giản hóa, ví dụ như xem xét, phê duyệt kết nối dữ liệu, chia sẻ dữ liệu, khai thác, phát triển dữ liệu. Các thủ tục hành chính này cần phải nhanh, thuận tiện để các cơ quan, tổ chức, doanh nghiệp thuận lợi trong việc tiếp cận, sử dụng dữ liệu, trong đó có dữ liệu dùng cho phát triển các giải pháp AI trong các cơ quan nhà nước.

Tận dụng các nguồn dữ liệu khác nhau: Để phát triển các giải pháp AI cho các cơ quan nhà nước, trong nhiều trường hợp cần kết nối, tận dụng dữ liệu do khu vực tư nắm giữ, sở hữu, chẳng hạn như dữ liệu về môi trường, thời tiết, đất đai, giao thông, thị trường lao động...

Đầu tư và tận dụng các hạ tầng AI: Cũng như cách tiếp cận về dữ liệu, đối với hạ tầng AI, trong ngắn hạn, có thể "liệu cơm gắp mắm", trong khuôn khổ có giới hạn của hạ tầng, nên phát triển các giải pháp AI đơn giản hơn, sử dụng ít năng lực tính toán như Tòa án Nhân dân Tối cao, tỉnh Tây Ninh (cũ) đã làm lâu nay. Hoặc điều chỉnh, giảm mức độ phức tạp của giải pháp AI cho phù hợp với dữ liệu và năng lực tính toán như trong trường hợp phát triển trợ lý ảo hỗ trợ rà soát văn bản quy phạm pháp luật.



Công cụ trí tuệ nhân tạo về pháp luật của Bộ Tư pháp.

Trong trung hạn và dài hạn, để nâng cấp các giải pháp AI hiện tại, hoặc phát triển các mô hình/giải pháp AI phức tạp trong các cơ quan nhà nước, ví dụ như cảnh báo cháy rừng, sạt lở đất, phân tích dữ liệu kinh tế - xã hội, cần dung lượng dữ liệu rất lớn, hạ tầng đám mây lưu trữ, hạ tầng tính toán cũng phải lớn.

Các cơ quan nhà nước có thể thiết lập mối quan hệ hợp tác để sử dụng các dịch vụ về: Hạ tầng trung tâm nền tảng điện toán đám mây, lưu trữ dữ liệu dùng cho AI... của các doanh nghiệp lớn như Viettel, FPT cho việc triển khai các ứng dụng AI. Trên thực tế, Viettel đã cung cấp hạ tầng cho một số địa phương, cụ thể như đang triển khai hạ tầng hỗ trợ tỉnh Hòa Bình (cũ), đặt ở Viettel; hỗ trợ Lạng Sơn xây dựng Trung tâm dữ liệu mới, chuyển từ Trung tâm dữ liệu cũ. Phương án thuê hạ tầng dịch vụ AI cũng được khuyến nghị cho các cơ quan nhà nước, vì tiết kiệm chi phí đầu tư, nâng cấp hệ thống, chi phí vận hành, bảo dưỡng thiết bị, không lãng phí tài nguyên.

Bên cạnh đó, phát triển nguồn năng lượng tái tạo như điện gió, điện mặt trời, cùng với triển khai các công nghệ lưu trữ năng lượng tiên tiến để duy trì nguồn điện ổn định, trong đó có điện cho vận hành các trung tâm dữ liệu. Đồng thời, cần áp dụng các công nghệ và biện pháp tiên tiến để giảm thiểu tiêu thụ năng lượng tại các trung tâm dữ liệu, chẳng hạn như hệ thống làm mát hiệu quả.