BỘ THÔNG TIN VÀ TRUYỀN THÔNG	CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
	Độc lập - Tự do - Hạnh phúc
Số:1418 /QĐ-BTTTT	*Hà Nội, ngày 31 tháng 07 năm 2023*

QUYẾT ĐỊNH

Ban hành kế hoạch thúc đẩy phát triển nền tảng công nghệ mô hình ngôn ngữ lớn tiếng Việt

BỘ TRƯỞNG BỘ THÔNG TIN VÀ TRUYỀN THÔNG

Căn cứ Nghị định số 48/2022/NĐ-CP ngày 26 tháng 7 năm 2022 của Chính phủ quy định chức năng, nhiệm vụ, quyền hạn và cơ cấu tổ chức của Bộ Thông tin và Truyền thông;

Căn cứ Nghị định số 73/2019/NĐ-CP ngày 05 tháng 9 năm 2019 của Chính phủ quy định quản lý đầu tư ứng dụng công nghệ thông tin sử dụng nguồn vốn ngân sách nhà nước;

Thực hiện Quyết định số 749/QĐ-TTg ngày 03 tháng 6 năm 2020 của Thủ tướng Chính phủ ban hành Chương trình Chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2030;

Thực hiện Quyết định số 942/QĐ-TTg ngày 15 tháng 6 năm 2021 của Thủ tướng Chính phủ phê duyệt Chiến lược phát triển Chính phủ điện tử hướng đến Chính phủ số giai đoạn 2021-2025, định hướng đến năm 2030;

Thực hiện Quyết định số 411/QĐ-TTg ngày Quyết định số 411/QĐ-TTg ngày 31 tháng 3 năm 2022 của Thủ tướng Chính phủ phê duyệt Chiến lược quốc gia phát triển kinh tế số và xã hội số đến năm 2025, định hướng đến năm 2030.

Theo đề nghị của Cục trưởng Cục Chuyển đổi số quốc gia.

QUYẾT ĐỊNH:

Điều 1. Ban hành kèm theo Quyết định này Kế hoạch Thúc đẩy phát triển nền tảng công nghệ mô hình ngôn ngữ lớn tiếng Việt.

Điều 2. Quyết định này có hiệu lực kể từ ngày ký.

Điều 3. Chánh Văn phòng, Cục trưởng Cục Chuyển đổi số quốc gia, Cục

trưởng Cục An toàn thông tin, Giám đốc Trung tâm Thông tin, Thủ trưởng các đơn vị thuộc, trực thuộc Bộ Thông tin và Truyền thông, tổ chức, cá nhân có liên quan chịu trách nhiệm thi hành Quyết định này./.

Nơi nhận:
- Như Điều 3;
- Bộ trưởng (để b/c);
- Các Thứ trưởng;
- Các cơ quan, đơn vị thuộc Bộ TT&TT;
- Lưu: VT, CĐSQG.

KT. BỘ TRƯỞNG
THỨ TRƯỞNG

BỘ THÔNG TIN VÀ TRUYỀN THÔNG

Nguyễn Huy Dũng

KẾ HOẠCH

Thúc đẩy phát triển nền tảng công nghệ mô hình ngôn ngữ lớn tiếng Việt

(Kèm theo Quyết định số /QĐ-BTTTT ngày / /2023 của Bộ Thông tin và Truyền thông)

1. Căn cứ thực hiện kế hoạch

1.1. Chương trình chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2023 phê duyệt tại Quyết định số 749/QĐ-TTg ngày 03/6/2020 đã xác định nhiệm vụ, giải pháp:

Lựa chọn ưu tiên nghiên cứu một số công nghệ cốt lõi mà Việt Nam có thể đi tắt đón đầu cũng như có khả năng tạo bứt phá mạnh mẽ như trí tuệ nhân tạo (AI), chuỗi khối (blockchain) và thực tế ảo/thực tế tăng cường (VR/AR).

Ưu đãi, hỗ trợ mạnh các doanh nghiệp khởi nghiệp phát triển và khuyến khích các doanh nghiệp lớn, truyền thống đi đầu trong việc ứng dụng các công nghệ này vào hoạt động sản xuất, thương mại.

Xây dựng hệ thống điện toán có năng lực đủ mạnh để xử lý, phân tích dữ liệu, huy động được sự tham gia của cộng đồng, cho phép các tổ chức, doanh nghiệp cùng khai thác phục vụ phát triển hệ sinh thái sản phẩm sáng tạo ứng dụng trí tuệ nhân tạo.

1.2. Chiến lược phát triển Chính phủ điện tử hướng tới Chính phủ số giai đoạn 2021-2025, định hướng đến 2030 phê duyệt tại Quyết định số 942/QĐ-TTg ngày 15/6/2021, Chiến lược quốc gia phát triển kinh tế số và xã hội số đến năm 2025, định hướng đến năm 2030 phê duyệt tại Quyết định số 411/QĐ-TTg ngày 31/3/2022 đã tiếp tục xác định nhiệm vụ, giải pháp: Xây dựng nền tảng trợ lý ảo để phục vụ cán bộ, công chức, viên chức và người dân.

2. Bối cảnh và sự cần thiết phát triển Mô hình ngôn ngữ lớn tiếng Việt

2.1. Mô hình ngôn ngữ lớn (LLM) có khả năng ứng dụng trong thực tế cuộc sống là một xu hướng mới nổi từ năm 2018. Sự xuất hiện chính thức của ChatGPT vào tháng 11/2022 cho thấy mô hình ngôn ngữ lớn là một cách tiếp cận mang tính đột phá trong việc phát triển trợ lý ảo có thể tương tác với con người theo cách tự nhiên. Mô hình ngôn ngữ lớn được phát triển bởi một số

hãng lớn trên thế giới mang tính đa ngôn ngữ, chưa chú trọng vào tiếng Việt, sử dụng dữ liệu được thu thập trên Internet, không hoàn toàn chính xác, lành mạnh, có thể dẫn đến những kết quả chưa tốt, thậm chí là kết quả sai.

Vì vậy, việc nghiên cứu, phát triển, đưa vào ứng dụng mô hình ngôn ngữ lớn tiếng Việt là một nhiệm vụ quan trọng, cần thiết và ý nghĩa. Mô hình ngôn ngữ lớn tiếng Việt sử dụng tri thức, dữ liệu huấn luyện đã được sàng lọc của Việt Nam, với chi phí thấp cho người dân, doanh nghiệp, tổ chức tại Việt Nam sử dụng để phát triển các ứng dụng mới.

2.2. Việc phát triển mô hình ngôn ngữ lớn tiếng Việt phải giải quyết hai thách thức chủ yếu. Một là thu thập, xử lý các nguồn dữ liệu tiếng Việt ở dạng thô để hình thành nên bộ dữ liệu đầy đủ tiếng Việt và bộ dữ liệu chỉ dẫn tiếng Việt (instruction dataset). Đây là vấn đề cơ bản nhất của việc phát triển một mô hình ngôn ngữ. Hai là thiết lập một hạ tầng tính toán cho phép thực hiện huấn luyện mô hình ngôn ngữ lớn.

Đây là vấn đề mới, vấn đề lớn của quốc gia. Bộ Thông tin và Truyền thông ban hành kế hoạch để thúc đẩy triển khai một cách thực chất, hiệu quả.

3. Mục tiêu đến năm 2025

- Việt Nam có ít nhất 01 nền tảng công nghệ mô hình ngôn ngữ lớn tiếng Việt, có khả năng cung cấp dịch vụ cho các nền tảng ứng dụng trí tuệ nhân tạo khác.

- 100% cơ quan nhà nước có trợ lý ảo giúp cán bộ, công chức phục vụ hoạt động của mình.

4. Các nhiệm vụ chính và tiến độ thực hiện

4.1. Xây dựng thể chế

a) Nội dung công việc:

Đề xuất xây dựng Luật Chính phủ số để tạo hành lang pháp lý đầy đủ cho phát triển chính phủ số; hoàn thiện các văn bản quy phạm pháp luật triển khai Luật Giao dịch điện tử (năm 2023); từ đó hoàn thiện thể chế giúp cán bộ công chức làm việc lên môi trường số, triển khai các hoạt động dựa trên công nghệ số.

b) Thời gian hoàn thành: Quý IV, năm 2025.

4.2. Xây dựng bộ dữ liệu lớn tiếng Việt để phát triển mô hình ngôn ngữ lớn

của Việt Nam

a) Nội dung công việc:

- Thu thập, tổng hợp, chuẩn hóa và cập nhật thường xuyên các dữ liệu, tri thức chung trên Internet;

- Thu thập, tổng hợp các dữ liệu, tri thức chung trong các tài liệu, ấn phẩm chính thức chưa được công bố rộng rãi trên Internet;

b) Thời gian hoàn thành:

- Hoàn thành bộ dữ liệu phục vụ thử nghiệm: Quý III/2023.

- Hoàn thành bộ dữ liệu lớn tiếng Việt để phát triển mô hình ngôn ngữ lớn của Việt Nam: Quý IV/2025.

4.3. Nghiên cứu, thử nghiệm và phát triển mô hình ngôn ngữ lớn Tiếng Việt

a) Nội dung công việc:

- Lựa chọn mô hình ngôn ngữ cơ sở, số lượng tham số phù hợp với quy mô bài toán ứng dụng đặt ra.

- Chuẩn bị hạ tầng tính toán để huấn luyện mô hình.

- Nền tảng dịch vụ mô hình ngôn ngữ lớn tiếng Việt với các thành phần cơ bản bao gồm công cụ phục vụ thu thập, xử lý, dán nhãn dữ liệu và các giao diện lập trình ứng dụng (API) phục vụ phát triển trợ lý ảo.

- Thực hiện huấn luyện mô hình ngôn ngữ tiếng Việt và đánh giá chất lượng mô hình.

b) Thời gian hoàn thành:

- Hoàn thành mô hình thử nghiệm: Quý IV/2023.

- Hoàn thành mô hình ngôn ngữ lớn tiếng Việt: Quý IV/2025.

4.4. Xây dựng Nền tảng trợ lý ảo cho cán bộ, công chức

a) Nội dung công việc:

- Phát triển Nền tảng Trợ lý ảo cho cán bộ, công chức;

- Chuẩn hóa quy trình và phát triển công cụ;

- Tích hợp LLM Tiếng Việt.

b) Thời gian hoàn thành:

- Hoàn thành thử nghiệm: Quý IV/2023.

- Hoàn thành nền tảng Trợ lý ảo cho cán bộ, công chức: Quý IV/2025.

5. Tổ chức thực hiện

5.1. Cục Chuyển đổi số quốc gia chủ trì, phối hợp với các đơn vị có liên quan triển khai cụ thể từng nội dung của kế hoạch theo đúng quy định hiện hành.

5.2. Các đơn vị thuộc, trực thuộc Bộ có trách nhiệm phối hợp thực hiện kế hoạch, xây dựng dữ liệu, triển khai sử dụng, đánh giá, nhận xét hoàn thiện sản phẩm./.