HƯỚNG DẪN YÊU CẦU CƠ BẢN VÀ PHƯƠNG PHÁP ĐÁNH GIÁ MÔ HÌNH NGÔN NGỮ LỚN TIẾNG VIỆT VÀ TRỢ LÝ ẢO VIỆT NAM (PHIÊN BẢN 1.0)

(*Ban hành kèm theo Văn bản số /BTTTT-CĐSQG ngày / /2024 của Bộ Thông tin và Truyền thông*)

DANH MỤC TỪ NGỮ VIẾT TẮT VÀ GIẢI THÍCH KHÁI NIỆM

Từ viết tắt	Nội dung đầy đủ/ khái niệm	Giải thích
AI	Artificial Intelligence	Trí tuệ nhân tạo.
LLM	Large Language Model	Mô hình ngôn ngữ lớn là một loại trí tuệ nhân tạo (AI) tiên tiến được thiết kế để xử lý, hiểu và tạo ra văn bản giống như con người. LLM được xây dựng và đào tạo trên một lượng lớn dữ liệu.
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên là một nhánh của trí tuệ nhân tạo (AI). NLP nghiên cứu sự tương tác giữa máy tính và ngôn ngữ của con người từ dạng văn bản đến tiếng nói.
NLU	Natural Language Understanding	Hiểu ngôn ngữ tự nhiên (NLU) là nhóm các kỹ thuật trong xử lý ngôn ngữ tự nhiên giúp hiểu ý định các câu hỏi của người dùng. NLU thông thường có hai thành phần chính là nhận diện ý định và nhận diện thực thể.
NLG	Natural Language Generation	Tạo ngôn ngữ tự nhiên (NLG) là nhóm các kỹ thuật trong xử lý ngôn ngữ tự nhiên giúp tự động tạo câu trả lời giống câu trả lời của con người.
Chatbot	Máy trò chuyện	Chatbot là chương trình máy tính được thiết kế để mô phỏng cuộc trò chuyện với người dùng thông qua giao diện văn bản hoặc giọng nói. Trước đây, chatbot dựa vào kỹ thuật xử lý ngôn ngữ tự nhiên (NLP, NLU, NLG) và học máy để hiểu và phản hồi đầu vào của người dùng. Sự phát triển của LLM cho phép Chatbot trả lời các câu hỏi về nhiều chủ đề. Kiến

Từ viết tắt	Nội dung đầy đủ/ khái niệm	Giải thích
		thức của chatbot thường giới hạn trong dữ liệu huấn luyện và thời gian giới hạn.
TLA	Trợ lý ảo AI	Trợ lý ảo là một Chatbot thông minh. Trợ lý ảo xử lý cả tương tác văn bản và giọng nói, tích hợp với nhiều thiết bị và dịch vụ để cung cấp trải nghiệm người dùng toàn diện. Trợ lý ảo có thể học từ tương tác của người dùng, thích ứng và cung cấp dịch vụ cá nhân hóa; có thể kết nối với nhiều nguồn dữ liệu và hệ thống, cung cấp trải nghiệm người dùng theo ngữ cảnh và hoạt động tốt trên các nền tảng khác nhau.
Web services	Dịch vụ Web	Là một cách để các ứng dụng hoặc hệ thống khác nhau có thể tương tác với nhau. Web services cho phép các ứng dụng hoặc máy chủ truyền dữ liệu và thông tin cho nhau thông qua mạng, bằng cách sử dụng các giao thức và tiêu chuẩn cụ thể.
CSDL	Cơ sở dữ liệu	
RAG	Retrieval Augmented Generation	Có nghĩa là Tạo tăng cường truy xuất. Đây là một kỹ thuật trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), kết hợp ưu điểm của các mô hình ngôn ngữ lớn (LLM) với hệ thống truy xuất thông tin.
RLHF	Reinforcement Learning from Human Feedback	Học tăng cường từ phản hồi của con người để cải thiện hiệu suất của mô hình.

I. MỤC ĐÍCH HƯỚNG DẪN

Căn cứ Quyết định số 749/QĐ-TTg ngày 03/6/2020 của Thủ tướng Chính phủ phê duyệt Chương trình Chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2030;

Căn cứ Quyết định số 942/QĐ-TTg ngày 15/06/2021 của Thủ tướng Chính phủ phê duyệt Chiến lược phát triển Chính phủ điện tử hướng tới Chính phủ số giai đoạn 2021 - 2025, định hướng đến năm 2030;

Căn cứ Quyết định số 411/QĐ-TTg ngày 31/03/2022 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;

Thực hiện Quyết định số 186/QĐ-BTTTT ngày 11/02/2022 của Bộ trưởng Bộ Thông tin và Truyền thông phê duyệt Chương trình thúc đẩy phát triển và sử dụng các nền tảng số quốc gia phục vụ chuyển đổi số, phát triển chính phủ số, kinh tế số, xã hội số; Quyết định số 2294/QĐ-BTTTT ngày 21/11/2023 của Bộ trưởng Bộ Thông tin và Truyền thông về việc sửa đổi, bổ sung Chương trình thúc đẩy phát triển và sử dụng các nền tảng số quốc gia phục vụ chuyển đổi số, phát triển chính phủ số, kinh tế số, xã hội số tại Quyết định số 186/QĐ-BTTTT ngày 11/02/2022.

Bộ Thông tin và Truyền thông hướng dẫn yêu cầu cơ bản và phương pháp đánh giá Mô hình ngôn ngữ lớn tiếng Việt và Trợ lý ảo Việt Nam (Phiên bản 1.0) (viết tắt là Tài liệu).

Các bộ, cơ quan ngang bộ, cơ quan thuộc Chính phủ, Ủy ban nhân dân các tỉnh, thành phố trực thuộc Trung ương tham khảo Tài liệu khi xây dựng, đánh giá, lựa chọn nền tảng Mô hình ngôn ngữ lớn tiếng Việt và Trợ lý ảo Việt Nam để triển khai phù hợp với nhu cầu thực tiễn. Các tổ chức, cá nhân có nhu cầu xây dựng, triển khai LLM tiếng Việt và TLA Việt Nam chủ động xem xét, áp dụng.

Tài liệu được xây dựng trên tinh thần mở, khuyến khích sự tham gia, bổ sung, hoàn thiện của các tổ chức, cá nhân và cộng đồng. Các ý kiến tham gia với Tài liệu, đề nghị gửi về Bộ Thông tin và Truyền thông (Cục Chuyển đổi số quốc gia) để nghiên cứu, hoàn thiện và ban hành các phiên bản tiếp theo.

II. NỘI DUNG HƯỚNG DẪN

1. Mục tiêu đánh giá LLM tiếng Việt và TLA Việt Nam

- Đánh giá khả năng hiểu câu hỏi và đưa ra câu trả lời phù hợp, chính xác;

- Đánh giá khả năng giao tiếp tự nhiên, hiểu và đáp ứng yêu cầu của người dùng một cách chính xác;

- Đánh giá khả năng tóm tắt chính xác nội dung chính của một đoạn văn;

- Đánh giá khả năng hiểu ngôn ngữ, dịch ngôn ngữ;

- Đánh giá khả năng viết văn bản sáng tạo.

2. Yêu cầu kỹ thuật đối với LLM tiếng Việt

a. Yêu cầu chung

- Đảm bảo có biện pháp kiểm soát và các thông tin, dữ liệu đầu ra của LLM tiếng Việt không vi phạm truyền thống lịch sử, văn hoá, đạo đức, địa lý, quan điểm chính trị, chủ quyền lãnh thổ, thuần phong mỹ tục của Việt Nam; không vi phạm các quy định về quyền tác giả, quyền sở hữu trí tuệ và các quyền liên quan;

- Đảm bảo phương thức cho phép các TLA Việt Nam và các ứng dụng khác có thể khai thác LLM tiếng Việt dưới dạng dịch vụ;

- Đảm bảo tuân thủ các quy định về an toàn thông tin mạng, an ninh mạng, bảo vệ dữ liệu cá nhân và các quy định khác có liên quan.

b. Yêu cầu cụ thể

STT	Yêu cầu	Mô tả chi tiết
1	Tạo nội dung bằng ngôn ngữ tự nhiên	- Có thể tạo nội dung bằng ngôn ngữ tự nhiên giống con người; - Văn bản tạo ra trôi chảy và mạch lạc; - Văn bản tạo ra chính xác về mặt ngữ pháp tiếng Việt.
2	Hiểu ngôn ngữ	- Có thể hiểu được ý định của câu hỏi bằng tiếng Việt; - Có thể hiểu được ý nghĩa của văn bản tiếng Việt; - Có thể phân tích ngữ pháp, ngữ nghĩa, nhận diện thực thể, gán nhãn ngữ nghĩa; - Có thể xác định, phân loại được các thông tin trong văn bản;
3	Có khả năng học hỏi	Có thể học hỏi từ dữ liệu mới để phục vụ các nhiệm vụ khác nhau.

STT	Yêu cầu	Mô tả chi tiết
4	Khả năng sáng tạo	- Có thể tạo ra các nội dung mới; - Có khả năng giải quyết các vấn đề một cách sáng tạo.
5	Khả năng giải quyết các yêu cầu cụ thể	Có khả năng thích ứng, giải quyết các yêu cầu của người dùng như tóm tắt văn bản tiếng Việt, phân tích văn bản,…
6	Khả năng giao tiếp	- Có thể giao tiếp hiệu quả đối với từng ngữ cảnh cụ thể. - Câu trả lời không chứa những nội dung xấu, độc, những vấn đề chính trị nhạy cảm.
7	Cung cấp dịch vụ để khai thác	Cung cấp các dịch vụ để các hệ thống TLA hoặc các hệ thống khác khai thác qua các hình thức như Webservices, API…

3. Yêu cầu kỹ thuật đối với TLA Việt Nam

a. Yêu cầu chung

- Đối với TLA triển khai theo kỹ thuật RAG dựa trên các LLM tiếng Việt, phải đảm bảo dữ liệu RAG không vi phạm truyền thống lịch sử, văn hoá, đạo đức, địa lý, quan điểm chính trị, chủ quyền lãnh thổ, thuần phong mỹ tục của Việt Nam; không vi phạm các quy định về quyền tác giả, quyền sở hữu trí tuệ và các quyền liên quan;

- Đảm bảo cung cấp phương thức để hỗ trợ người dùng thực hiện kỹ thuật RLHF;

- Cho phép lựa chọn khai thác các LLM tiếng Việt khác nhau;

- Đảm bảo tuân thủ các quy định về an toàn thông tin mạng, an ninh mạng, bảo vệ dữ liệu cá nhân và các quy định khác có liên quan.

b. Yêu cầu cụ thể

STT	Yêu cầu	Mô tả chi tiết
	Quản lý tài khoản	

STT	Yêu cầu	Mô tả chi tiết
1	Quản lý tài khoản sử dụng hệ thống	Bao gồm các yêu cầu sau: - Quản lý danh sách tài khoản hệ thống; - Có thể thêm, sửa, xóa tài khoản; - Phân quyền tài khoản, gán tài khoản vào nhóm quyền.
2	Quản lý nhóm quyền	- Có thể thêm, sửa, xóa nhóm quyền; - Có thể thêm, sửa, xóa người dùng theo nhóm quyền; - Có thể phân quyền theo nhóm quyền.
3	Đăng nhập	Có thể đăng nhập hệ thống.
4	Đăng xuất	Có thể đăng xuất hệ thống.
	Quản lý trò chuyện	
5	Tạo cuộc trò chuyện mới	- Có thể tạo một cuộc trò chuyện mới; - Có thể gợi ý theo chủ đề, theo thói quen…; - Có thể hiểu và lưu ngữ cảnh cuộc trò truyện; - Hiển thị danh sách các cuộc trò chuyện gần nhất.
6	Thực hiện câu hỏi	- Cho phép đặt câu hỏi dưới dạng văn bản (text); - Cho phép đặt câu hỏi dưới dạng giọng nói.
7	Trả lời câu hỏi	- Cho phép hiển thị dưới các dạng cơ bản sau: text, ảnh, bảng, biểu … - Cho phép hiển thị các trích dẫn, các đường dẫn đến nguồn thông tin truy xuất (nếu có); - Có thể tương tác với câu trả lời như: + Thích, không thích; + Góp ý câu trả lời đúng.

STT	Yêu cầu	Mô tả chi tiết
		- Chất lượng câu trả lời tốt. Việc đánh giá chất lượng câu trả lời được thực hiện theo Mô hình đánh giá (*Mô tả chi tiết tại Mục 4).*
8	Chia sẻ dữ liệu	Cho phép chia sẻ dữ liệu bằng một trong các hình thức sau: đường dẫn truy cập, xuất ra tệp…
9	Báo cáo vi phạm	Cho phép báo cáo vi phạm nếu nội dung câu trả lời vi phạm truyền thống lịch sử, văn hoá, đạo đức, địa lý, quan điểm chính trị, chủ quyền lãnh thổ, thuần phong mỹ tục của Việt Nam hoặc các quy định hiện hành.
10	Xem lịch sử trò chuyện	- Người dùng có thể xem danh sách lịch sử trò chuyện; - Người dùng có thể xóa lịch sử trò chuyện.
	Các yêu cầu khác	
11	Trợ giúp	- Có hướng dẫn sử dụng; - Thông tin phiên bản; - Hiển thị câu hỏi thường gặp.
12	Báo cáo, thống kê	Xem báo cáo, thống kê: - Thống kê theo người dùng; - Thống kê theo câu hỏi; - Thống kê theo câu trả lời; - Thống kê theo thời gian.
13	Đào tạo TLA	- Có công cụ để đào tạo TLA; - TLA có thể trả lời tối thiểu theo 03 phương pháp sau: + Có thể trả lời các câu hỏi mang tính quy trình gồm nhiều bước; + Có thể trả lời dựa trên ngân hàng các câu hỏi-câu trả lời;

STT	Yêu cầu	Mô tả chi tiết
		+ Có thể trả lời dựa trên LLM.
14	Tích hợp với các phần mềm, hệ thống, nền tảng khác	Có khả năng tích hợp với các phần mềm, hệ thống, nền tảng khác để khai thác, sử dụng.

4. Phương pháp đánh giá

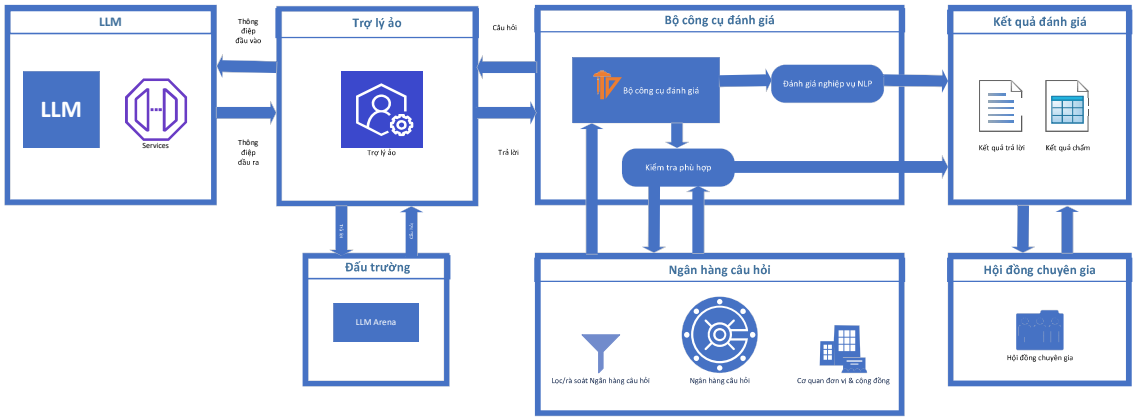
Việc đánh giá LLM tiếng Việt và TLA Việt Nam được thực hiện theo mô hình sau:

LLM tiếng Việt hoặc TLA Việt Nam là đối tượng được đánh giá.

Việc đánh giá LLM tiếng Việt hoặc TLA Việt Nam được thực hiện thông qua các câu hỏi-đáp lấy từ từ **Ngân hàng câu hỏi**. Việc lựa chọn các câu hỏi-đáp tuỳ thuộc vào mỗi **Bài đánh giá**. *Ví dụ: để đánh giá LLM tiếng Việt hoặc TLA Việt Nam phục vụ cho mục đích hỏi đáp chung, hỏi đáp tổng hợp nhiều vấn đề thì lựa chọn đầy đủ câu hỏi-đáp theo các nhóm và các lĩnh vực khác nhau; để đánh giá TLA Việt Nam phục vụ cho các chuyên ngành, có thể không cần lựa chọn các câu hỏi-đáp đa dạng thuộc nhiều nhóm, nhiều lĩnh vực, chỉ cần lựa chọn các câu hỏi-đáp trong lĩnh vực chuyên ngành.*

Cơ quan, tổ chức khi áp dụng Tài liệu này, có thể lựa chọn hoặc xây dựng các **Bài đánh giá** phù hợp với mục tiêu sử dụng của LLM tiếng Việt hoặc TLA Việt Nam.

Kết quả đánh giá có thể được thực hiện bằng công cụ tự động, gọi chung là **Bộ công cụ đánh giá** và thực hiện bằng con người đánh giá. Tuy nhiên, kết luận cuối cùng là của con người, gọi chung là **Hội đồng chuyên gia**.



Mô tả cụ thể mô hình:

(1) **LLM**: là LLM tiếng Việt được đánh giá. Việc đánh giá này không thực hiện trực tiếp với LLM tiếng Việt mà thực hiện thông qua một ứng dụng/ giao diện cho phép đưa vào các câu hỏi và nhận được các câu trả lời (hoặc phản hồi). Các ứng dụng/ giao diện đó gọi chung là Trợ lý ảo.

(2) **Trợ lý ảo**: là TLA Việt Nam được đánh giá, hoặc ứng dụng/ giao diện dùng để đánh giá LLM tiếng Việt (trong trường hợp đánh giá LLM tiếng Việt).

(3) **Bộ công cụ đánh giá**: là Bộ công cụ đánh giá do Bộ Thông tin và Truyền thông thiết lập trên cơ sở tham gia của cộng đồng. Bao gồm các công cụ cho phép thực hiện tự động các tác vụ nhằm đánh giá khả năng của LLM tiếng Việt hoặc TLA Việt Nam. Trong đó có 3 công cụ chính:

- Công cụ tự động lấy ngẫu nhiên các câu hỏi trong **Ngân hàng câu hỏi** để đưa vào **Trợ lý ảo** và nhận lại câu trả lời (hoặc phản hồi);

- Công cụ tự động kiểm tra phù hợp giữa câu trả lời của **Trợ lý ảo** so với câu trả lời có sẵn trong **Ngân hàng câu hỏi**;

- Công cụ đánh giá nghiệp vụ NLP, NLU, NLG bằng cách tự động đưa ra các yêu cầu cho **Trợ lý ảo** và nhận về câu trả lời (hoặc phản hồi). Ví dụ: yêu cầu tóm tắt một đoạn văn, yêu cầu làm một bài thơ, yêu cầu phiên dịch…

Bộ Thông tin và Truyền thông chủ động việc xây dựng các công cụ tự động nói trên và khuyến khích các tổ chức, cá nhân tham gia đóng góp để hoàn thiện Bộ công cụ đánh giá.

(4) **Ngân hàng câu hỏi**: chứa các câu hỏi và câu trả lời do Bộ Thông tin và Truyền thông phát triển trên cơ sở tham gia đóng góp của cộng đồng. Câu hỏi và câu trả lời được được phân chia thành các lĩnh vực khác nhau nhằm đánh giá được mức độ thông hiểu tiếng Việt và mức độ hiệu quả trong việc giải quyết các bài toán nghiệp vụ chuyên ngành cụ thể.

Ngân hàng câu hỏi bao gồm các nhóm như sau:

STT	Tên nhóm	Giải thích
1	Câu hỏi mở	Dạng câu hỏi này nhằm đánh giá khả năng hiểu và trả lời các câu hỏi phức tạp, đòi hỏi sự suy luận và sáng tạo của LLM tiếng Việt và TLA Việt Nam.

2	Câu hỏi trắc nghiệm	Dạng câu hỏi này cung cấp một số lựa chọn trả lời, nhằm đánh giá khả năng trả lời chính xác của LLM tiếng Việt và TLA Việt Nam.
3	Câu hỏi yêu cầu thực hiện một tác vụ	Dạng câu hỏi này yêu cầu LLM tiếng Việt và TLA Việt Nam thực hiện một nhiệm vụ cụ thể, chẳng hạn như viết một bài thơ, dịch một đoạn văn bản hoặc tóm tắt một bài báo.
4	Câu hỏi yêu cầu cung cấp thông tin	Dạng câu hỏi này yêu cầu LLM tiếng Việt và TLA Việt Nam cung cấp thông tin về một chủ đề cụ thể.

Ngân hàng câu hỏi bao gồm các lĩnh vực:

STT	Tên lĩnh vực
I	**STEM**
1	Toán tiểu học
2	Toán trung học cơ sở
3	Toán trung phổ thông
4	Vật lý trung học cơ sở
5	Vật lý trung học phổ thông
6	Hóa học trung học cơ sở
7	Hóa học trung học phổ thông
8	Sinh học trung học cơ sở
9	Sinh học trung học phổ thông
10	Tin học ứng dụng
11	Kiến trúc máy tính
13	Mạng máy tính

STT	Tên lĩnh vực
14	Toán rời rạc
15	Kỹ thuật điện
16	Giới thiệu về lập trình
17	Kỹ sư đo lường
18	Hệ điều hành
19	Thống kê và xác suất
II	**Khoa học xã hội**
20	Xã hội học
21	Nhân khẩu học
22	Tâm lý học
23	Kinh tế học
24	Chính trị học
25	Tội phạm học
26	Địa lý
27	Giáo dục
28	Nghiên cứu môi trường
29	Thông tin đại chúng và truyền thông
30	Hành chính công
31	Chủ nghĩa Mác – Lênin
32	Tư tưởng Hồ Chí Minh
33	Nền tảng tư tưởng, tư liệu, văn kiện của Đảng Cộng sản Việt Nam

STT	Tên lĩnh vực
34	Giáo dục công dân
35	Nghiên cứu giao tiếp
36	Quản trị kinh doanh
37	Kinh tế vĩ mô
38	Kinh tế vi mô
III	**Nhân văn**
39	Ngôn ngữ học
40	Lịch sử
41	Luật pháp
42	Văn học
43	Nghệ thuật
44	Triết học
45	Tôn giáo
IV	**Lĩnh vực khác**
46	Văn hoá - thể thao
47	Môi trường
48	Danh lam, thắng cảnh, du lịch
49	Giải trí
50	Nghề nghiệp
IV	**Lĩnh vực chuyên ngành: Gồm các lĩnh vực chuyên ngành (trong trường hợp đánh giá các TLA Việt Nam phục vụ các chuyên ngành riêng)**

Ngân hàng câu hỏi được Bộ Thông tin và Truyền thông cùng cộng đồng xây dựng mang tính chất mở. Khuyến khích các tổ chức, cá nhân bổ sung, phát triển thêm các câu hỏi và câu trả lời.

Ngân hàng câu hỏi do Cục Chuyển đổi số quốc gia quản lý và cập nhật. Các tổ chức, cá nhân có đóng góp câu hỏi – đáp cho Ngân hàng câu hỏi đề nghị liên hệ với Cục Chuyển đổi số quốc gia.

Việc áp dụng toàn bộ hay một số lĩnh vực, nhóm lĩnh vực để đánh giá LLM tiếng Việt và TLA Việt Nam do Hội đồng chuyên gia quyết định.

(5) **Kết quả đánh giá**: Là kết quả việc đánh giá bằng **Bộ công cụ đánh giá**. Bao gồm Kết quả trả lời (hoặc phản hồi) của **Trợ lý ảo** và Kết quả chấm điểm của **Bộ công cụ đánh giá**.

Việc chấm điểm được thực hiện theo trình tự như sau:

Bước 1: **Bộ công cụ đánh giá** chọn câu hỏi ngẫu nhiên từ **ngân hàng câu hỏi** để đưa vào **Trợ lý ảo**. Số lượng câu hỏi được phân bố đều theo từng nhóm và lĩnh vực, đảm bảo phù hợp với **Bài đánh giá** (*một số bài đánh giá điển hình tại mục 5*). Mỗi **Bài đánh giá** nên xác định rõ số lượng câu hỏi, lựa chọn các câu hỏi từ dễ đến khó và xác định cơ cấu điểm cho mỗi câu trả lời đúng. Nên sử dụng thang điểm tối đa là 100 điểm hoặc 100 phần trăm (%).

Bước 2: **Bộ công cụ đánh giá** nhận lại câu trả lời (hoặc phản hồi) của **Trợ lý ảo** với từng câu hỏi và kiểm tra phù hợp;

Bước 3: **Bộ công cụ đánh giá** sẽ tính số điểm đạt được theo từng lĩnh vực và từng nhóm câu hỏi; điểm số cuối cùng là điểm trung bình theo từng lĩnh vực, từng nhóm câu hỏi nói trên.

Kết quả này mang tính chất tham khảo, sẽ được **Hội đồng chuyên gia** đánh giá và đưa ra kết luận cuối cùng.

(6) **Hội đồng chuyên gia**:

Để đánh giá LLM tiếng Việt và TLA Việt Nam thuộc Chương trình thúc đẩy phát triển và sử dụng các nền tảng số quốc gia phục vụ chuyển đổi số, phát triển chính phủ số, kinh tế số, xã hội số, Bộ Thông tin và Truyền thông thành lập Hội đồng chuyên gia, mời các cá nhân có kinh nghiệm trong nghiên cứu, phát triển và triển khai LLM tiếng Việt và TLA Việt Nam tham gia.

Các tổ chức, cá nhân khi áp dụng Tài liệu đánh giá này có thể chủ động thành lập các Hội đồng chuyên gia phù hợp.

Hội đồng chuyên gia có trách nhiệm đánh giá và đưa ra kết luận cuối cùng đối với mỗi LLM tiếng Việt và TLA Việt Nam.

(7) **Đấu trường (LLM Arena)**: là môi trường để các LLM tiếng Việt và TLA Việt Nam đăng ký, công khai cho cộng đồng đánh giá. Kết quả do cộng đồng đánh giá được công bố công khai, minh bạch. Đấu trường được Bộ Thông tin và Truyền thông thiết lập. Kết quả, bảng xếp hạng các LLM tiếng Việt và TLA Việt Nam được công bố trên Đấu trường là cơ sở để các tổ chức, cá nhân tham khảo, lựa chọn triển khai thử nghiệm.

5. Một số bài đánh giá điển hình

Tùy vào yêu cầu cụ thể, Hội đồng chuyên gia có thể lựa chọn áp dụng các bài đánh giá phù hợp. Sau đây là một số bài đánh giá điển hình có thể tham khảo:

- Needle in a haystack (NIAH): đây là bài đánh giá chuyên cho LLM được tăng cường truy xuất (LLM RAG);

- Massive Multitask Language Understanding (MMLU): Đánh giá về khả năng hiểu ngôn ngữ để thực hiện nhiệm vụ trên nhiều miền, lĩnh vực khác nhau;

- LogiQA: là bài đánh giá để kiểm tra khả năng suy luận của LLM và Trợ lý ảo;

- GSM1k: là bài đánh giá để đo lường mức độ khớp và khả năng suy luận trong các LLM;

- C-Eval là một bài đánh giá toàn diện bao gồm nhiều chuyên ngành và mức độ khó khác nhau dành cho các mô hình LLM.

- Vibe-Eval: Là một tiêu chuẩn đánh giá tiên tiến cho các mô hình ngôn ngữ đa phương thức AI, nổi bật với khung đánh giá có cấu trúc, kiểm tra chặt chẽ khả năng hiểu biết trực quan của các mô hình;

- ViLLM-Eval: là bộ đánh giá toàn diện đầu tiên được thiết kế riêng để đo lường kiến thức và khả năng lập luận của các LLM trong tiếng Việt, bao gồm các câu hỏi trắc nghiệm và tác vụ dự đoán từ tiếp theo ở nhiều cấp độ khó, đa dạng lĩnh vực từ khoa học nhân văn đến khoa học kỹ thuật

Các bài đánh giá điển hình được trình bày tóm tắt tại Phụ lục kèm theo./