

Xây Dựng Ứng Dụng RAG Cho Luật Đất Đai

MÔN HỌC: PHÂN TÍCH VĂN BẢN VÀ ỨNG DỤNG

GIẢNG VIÊN HƯỚNG DẪN: TS. Ngô Minh Mẫn

THÀNH VIÊN NHÓM: 22C01016 - Tống Lương Hương Quỳnh
22C01026 - Nguyễn Ngọc Thảo Uyên
22C01034 - Đào Thị Minh Lý

Link Colab:  app.ipynb

Link data:  data

MỤC LỤC

1. Giới thiệu về RAG.....	3
1.1. RAG Workflow.....	4
1.2. Ứng dụng của RAG trong LLM.....	4
1.3. RAG với Llama Index.....	5
2. Ứng dụng RAG cho Luật Đất Đai.....	6
2.1. Hướng tiếp cận VinaLlama + mpnet (multilingual embedding).....	6
a. Mô hình VinaLlama.....	6
b. Mô hình mpnet.....	7
c. Mô hình RAG với Vinallama.....	7
2.2. Hướng tiếp cận Llama2 + NLLB (translation model).....	8
a. Mô hình Llama2.....	8
b. Mô hình NLLB.....	8
c. Mô hình RAG với Llama2.....	9
3. So sánh hai hướng tiếp cận.....	9
3.1. Test case 1.....	10
3.2. Test case 2.....	11
3.3. Test case 3.....	13
3.4. Test case 4.....	14
3.5. Nhận xét.....	15
4. Demo.....	15
4.1. Gradio.....	15
4.2. Giao diện.....	16
4.2.1. DEMO 1.....	16
4.2.2. DEMO 2.....	16
4.3. Cách sử dụng.....	17
REFERENCES.....	18

1. Giới thiệu về RAG

Gần đây, sự phát triển của các mô hình ngôn ngữ lớn (LLMs) như ChatGPT, LLama-2, Mistral,... đã mở đường cho những tiến bộ vượt bậc trong lĩnh vực xử lý ngôn ngữ tự nhiên cũng như đẩy mạnh việc ứng dụng AI vào cuộc sống. Tuy nhiên, những mô hình mạnh mẽ này cũng đi kèm với một số thách thức cần phải giải quyết. Một trong những vấn đề lớn là hiện tượng "hallucination" - ảo giác, tức việc LLM tạo ra các thông tin không chính xác, không đúng sự thật hoặc không được hỗ trợ bởi dữ liệu có sẵn. Hiện tượng này rất nguy hiểm bởi nó có thể dẫn đến việc cung cấp các thông tin sai lệch gây hậu quả nghiêm trọng và làm giảm độ tin cậy của các hệ thống dựa trên AI. Để giải quyết các vấn đề này, một phương pháp thay thế được gọi là Retrieval-Augmented Generation (RAG) được phát triển. Trong RAG, dữ liệu của người dùng được tải và chuẩn bị cho các truy vấn hoặc được lập chỉ mục (indexed). Truy vấn của người dùng hoạt động dựa trên chỉ mục, giúp lọc dữ liệu của người dùng theo ngữ cảnh phù hợp nhất. Sau đó, ngữ cảnh này và truy vấn của người dùng sẽ chuyển đến LLM cùng với lời nhắc và LLM sẽ cung cấp phản hồi.

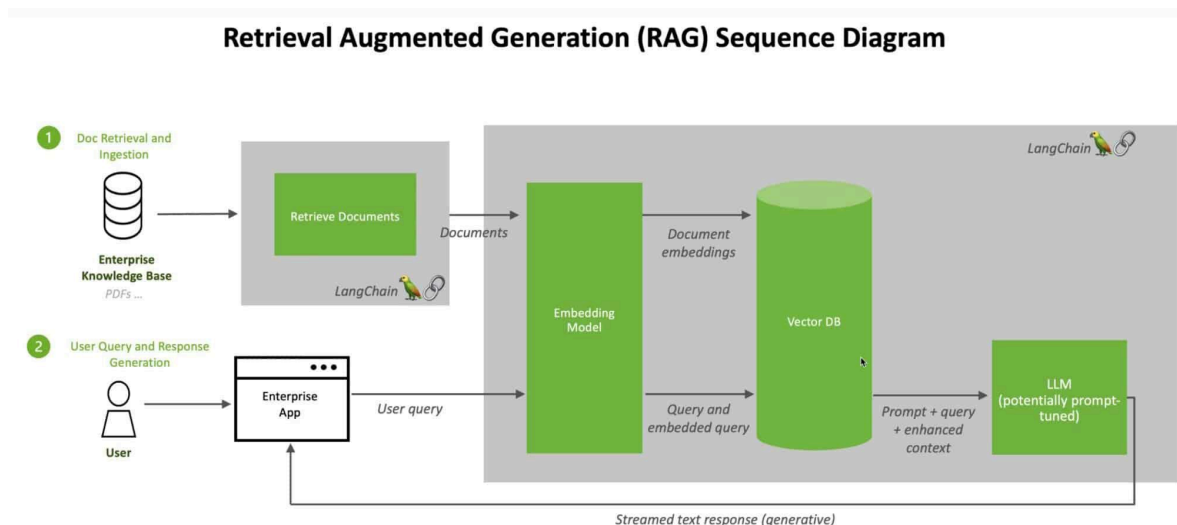
RAG là một kỹ thuật mạnh mẽ trong lĩnh vực xử lý ngôn ngữ tự nhiên, kết hợp hai khả năng quan trọng trong quá trình tạo nội dung của các mô hình LLM: khả năng "Retrieval" (truy xuất thông tin) và khả năng "Generation" (tạo sinh văn bản). Trình truy xuất tài liệu chịu trách nhiệm tìm kiếm thông tin liên quan từ một số lượng lớn tài liệu dựa trên truy vấn đầu vào. Thông tin này sau đó được chuyển đến LLM, để tạo ra phản hồi. RAG giúp cải thiện tính chính xác và độ tin cậy của các câu trả lời được tạo ra bởi LLM.

Có năm giai đoạn chính trong RAG:

- Loading: đề cập đến việc đưa dữ liệu của người dùng vào mô hình để thực hiện truy vấn. Dữ liệu có thể là text, file pdf, link, dữ liệu truy vấn qua database.
- Indexing: điều này có nghĩa là tạo cấu trúc dữ liệu cho phép truy vấn dữ liệu. Đối với LLM, điều này gần như luôn có nghĩa là tạo các vector nhúng, biểu diễn bằng số về ý nghĩa dữ liệu của người dùng cũng như nhiều chiến lược siêu dữ liệu khác để giúp người dùng dễ dàng tìm thấy dữ liệu chính xác có liên quan theo ngữ cảnh.
- Storing: sau khi dữ liệu của người dùng được lập chỉ mục, người dùng sẽ luôn lưu trữ chỉ mục của mình cũng như siêu dữ liệu khác.
- Querying: có thể sử dụng cấu trúc dữ liệu LLM và Llama Index để truy vấn.
- Evaluation: là kiểm tra mức độ hiệu quả của nó so với các chiến lược khác hoặc khi người dùng thực hiện thay đổi. Đánh giá cung cấp các thước đo khách quan về mức độ chính xác, trung thực và nhanh chóng của phản hồi của người dùng đối với các truy vấn.

RAG dùng LLM cho phép mô hình tận dụng kiến thức ngoài từ một kho dữ liệu văn bản lớn, giúp câu trả lời chính xác và mang tính thông tin hơn. Ngoài ra, RAG còn giúp giảm bớt sự hallucination trong câu trả lời và tiết kiệm chi phí.

1.1. RAG Workflow



Khi người dùng đặt câu hỏi, mô hình sẽ gửi truy vấn đến mô hình embedding và chuyển câu query thành dạng vector số

Sau đó, mô hình nhúng sẽ so sánh vector số này với các vector trong cơ sở dữ liệu vector

Khi tìm thấy một hoặc nhiều kết quả trùng khớp, nó sẽ truy xuất dữ liệu liên quan, chuyển đổi nó thành các từ mà con người có thể đọc được và chuyển lại cho LLM.

Cuối cùng, LLM kết hợp các từ được truy xuất và phản hồi của chính nó đối với truy vấn thành câu trả lời cuối cùng mà nó đưa ra cho người dùng, có khả năng trích dẫn các nguồn mà mô hình nhúng tìm thấy.

1.2. Ứng dụng của RAG trong LLM

RAG kết hợp với LLMs có nhiều ứng dụng trong lĩnh vực AI và xử lý ngôn ngữ tự nhiên (NLP) như:

- Trả lời câu hỏi: RAG tốt trong việc trả lời câu hỏi, nhờ khả năng truy xuất tài liệu và tạo câu trả lời chính xác, phù hợp với ngữ cảnh.
- Tóm tắt văn bản: RAG có thể dùng để tạo tóm tắt cho các tài liệu dài, thông qua việc tìm ra những phần quan trọng nhất của tài liệu.
- Hoàn thành văn bản: Mô hình RAG có thể dùng để hoàn thành các đoạn văn bản dựa trên nội dung và ngữ cảnh hiện có, như viết thư điện tử hoặc hoàn thiện mã nguồn.
- Dịch: Dù không phải là mục đích chính, RAG cũng có thể được sử dụng cho nhiệm vụ dịch thuật, thông qua việc truy xuất các bản dịch liên quan từ dữ liệu và tạo ra bản dịch phù hợp với những ví dụ đã có.

1.3. RAG với Llama Index

Llama Index là một khung điều phối hoặc “khung dữ liệu” giúp đơn giản hóa việc xây dựng các ứng dụng LLM. Nó cung cấp khả năng thực hiện tăng cường dữ liệu của dữ liệu riêng tư, cho phép nó được tích hợp vào LLM để tạo và lý luận kiến thức. Trọng tâm của tất cả chức

năng AI tổng quát là dữ liệu. Các ứng dụng doanh nghiệp cần có khả năng truy cập nhiều hơn chỉ là dữ liệu công khai mà LLM được đào tạo và cần kết hợp dữ liệu có cấu trúc, phi cấu trúc và bán cấu trúc từ tất cả các nguồn dữ liệu bên trong và bên ngoài của chúng để xây dựng ứng dụng.

Chính sự tích hợp dữ liệu này mà Llama Index cung cấp, đưa dữ liệu từ nhiều nguồn duy nhất, nhúng dữ liệu đó dưới dạng vector, lưu trữ dữ liệu mới được vector hóa trong cơ sở dữ liệu vector. Cuối cùng, cho phép các ứng dụng sử dụng dữ liệu đó để thực hiện các hoạt động phức tạp với thời gian phản hồi có độ trễ thấp như tìm kiếm vector.

Llama Index sử dụng mô hình Retrieval-Augmented Generation (RAG) để tránh chi phí huấn luyện và cập nhật. Thay vì tinh chỉnh LLM, RAG kết hợp thông tin từ nguồn dữ liệu của người dùng vào câu hỏi và yêu cầu LLM trả lời dựa trên ngữ cảnh được bổ sung. Điều này giúp vượt qua ba hạn chế của phương pháp tinh chỉnh: không cần huấn luyện, luôn cập nhật và có thể xem tài liệu đã truy xuất.

Các bước thực hiện xây dựng công cụ RAG với Llama Index như sau:

- **Tạo môi trường và cài đặt thư viện**

Cài đặt môi trường và thư viện cho RAG với Llama Index. Tải mô hình LLM ví dụ như gpt-3.5-turbo, llama2,...

- **Load document**

Để trợ giúp LLM, chúng tôi cần cung cấp cho LLM thông tin liên quan từ các nguồn kiến thức. Các nguồn kiến thức này có thể là dữ liệu có cấu trúc như CSV, Bảng tính hoặc bảng SQL, dữ liệu phi cấu trúc như văn bản, Tài liệu Word, Google Tài liệu, PDF hoặc PPT và dữ liệu bán cấu trúc như Notion, Slack, Salesforce, v.v. Trong bài viết này, chúng tôi sẽ sử dụng tệp PDF. Llama Index bao gồm một lớp SimpleDirectoryReader, lớp này có thể đọc các tài liệu đã lưu từ một thư mục được chỉ định. Nó tự động chọn một trình phân tích cú pháp dựa trên phần mở rộng tệp tin.

```
documents = SimpleDirectoryReader(input_dir='data').load_data()
```

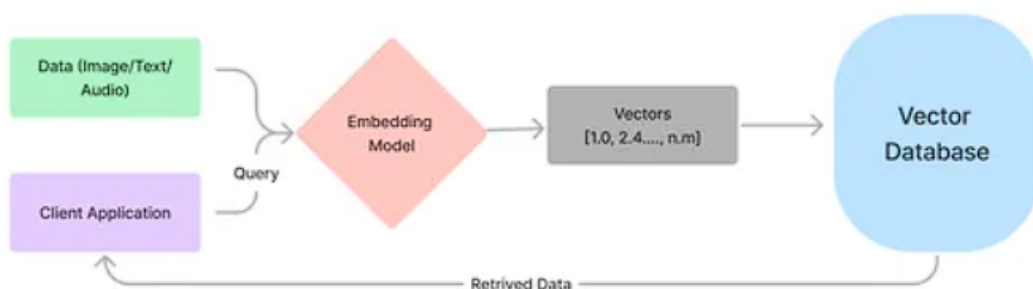
- **Tạo các đoạn văn bản (text chunks)**

Thông thường, dữ liệu được trích xuất từ các nguồn kiến thức rất dài, vượt quá phạm vi ngữ cảnh của LLM. Một cách để giải quyết vấn đề này là phân đoạn văn bản. Trong phân đoạn văn bản, văn bản dài hơn được chia thành các phần nhỏ hơn dựa trên dấu phân cách. Các đoạn văn bản nhỏ hơn mang lại độ chính xác khi nhúng tốt hơn, sau đó cải thiện độ chính xác khi truy xuất. SimpleNodeParser tạo các nút từ các đoạn văn bản và các đoạn văn bản được tạo bằng cách sử dụng TokenTextSplitter của Llama Index. Chúng ta cũng có thể sử dụng SentenceSplitter.

- **Xây dựng cơ sở kiến thức (Knowledge Bases)**

Để lưu trữ và vận hành dữ liệu trong RAG, chúng ta cần cơ sở dữ liệu vector. Cơ sở dữ liệu vector là kho lưu trữ dữ liệu được xây dựng có mục đích để lưu trữ và truy vấn vector.

- Embeddings: Hai văn bản có liên quan về mặt ngữ nghĩa sẽ ở gần nhau trong không gian vector, trong khi các văn bản khác nhau sẽ ở xa nhau. Phần nhúng có khả năng phi thường trong việc ánh xạ sự tương tự giữa các điểm dữ liệu khác nhau. Khi người dùng gửi một truy vấn văn bản, chúng tôi sẽ chuyển đổi nó thành các phần nhúng bằng cách sử dụng cùng một mô hình, so sánh khoảng cách của các phần nhúng văn bản được lưu trữ trong cơ sở dữ liệu vector và truy xuất các đoạn văn bản “n” gần nhất. Những đoạn này là những đoạn có ngữ nghĩa giống nhất với văn bản được truy vấn. Llama Index có triển khai tùy chỉnh các mô hình nhúng phổ biến.
- Cơ sở dữ liệu vector: Cơ sở dữ liệu vector được xây dựng có mục đích để lưu trữ và sắp xếp các phần nhúng cũng như siêu dữ liệu liên quan nhằm mang lại hiệu quả truy vấn tối đa. Chúng cung cấp khả năng truy xuất dữ liệu theo ngữ nghĩa, giúp tăng cường LLM bằng thông tin mới.



- **Truy vấn**

Bước cuối cùng là truy vấn từ chỉ mục và nhận phản hồi từ LLM.

2. Ứng dụng RAG cho Luật Đất Đai

2.1. Hướng tiếp cận VinaLlama + mpnet (multilingual embedding)

a. Mô hình VinaLlama

Mô hình VinaLlama là một mô hình ngôn ngữ lớn (LLM - Large Language Model) được phát triển bởi nhóm kỹ sư GenZ và VILM (Vietnamese Institute of Language and Media). Được xây dựng trên nền tảng của LLaMA-2 với thêm 800 tỷ token đào tạo bổ sung, VinaLlama là một mô hình tiếng Việt có khả năng xử lý ngôn ngữ tự nhiên mạnh mẽ. Mô hình này hiện đã truy cập được trên HuggingFace, đảm bảo tính tương thích với tất cả các thư viện được hỗ trợ bởi 'Transformers'.

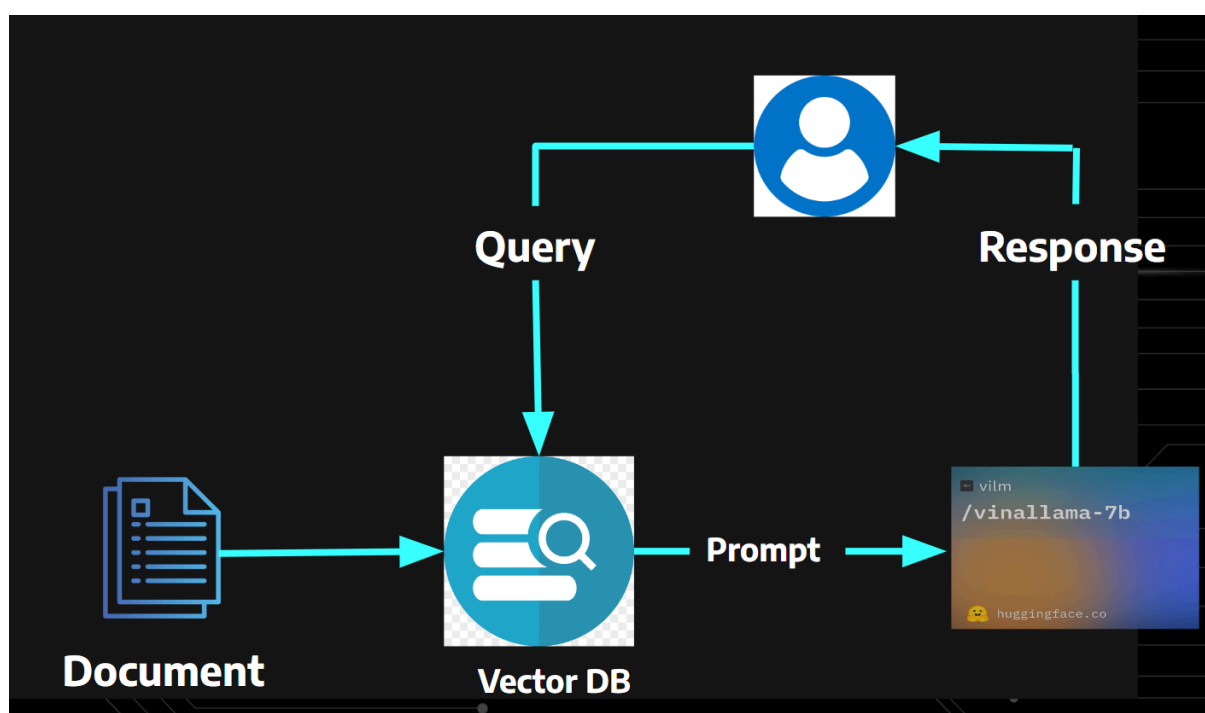
Lựa Chọn Tokenizer: VinaLlama sử dụng tokenizer LLaMA-2-chat của BKAI [5] đặc biệt được tạo ra để xử lý tiếng Việt. Tokenizer này đã được chứng minh hiệu suất cao trong việc xử lý ngôn ngữ Việt.

VinaLLaMA-7B: Có quy mô tương đương, ngoạn mục vượt trội so với các mô hình lớn hơn, kể cả những mô hình có 13 tỷ tham số, về điểm trung bình. Thành công này là minh chứng cho hiệu quả của quá trình tinh chỉnh và khả năng linh hoạt của mô hình trong việc tận dụng quá trình đào tạo để đạt hiệu suất xuất sắc, ngay cả so với các đối thủ lớn hơn.

b. Mô hình MPNet

MPNet là một mô hình nhúng văn bản đa ngôn ngữ (Multilingual Embedding) mạnh mẽ, mang lại khả năng biểu diễn chất lượng cao cho văn bản trong nhiều ngôn ngữ khác nhau. Mô hình này kết hợp masked language modeling (MLM) và permuted language modeling (PLM) trong một góc nhìn, tận dụng ưu điểm của cả hai phương pháp để tạo ra biểu diễn văn bản đa ngôn ngữ linh hoạt. MPNet được áp dụng để biểu diễn từng từ trong văn bản dưới dạng vector. Điều này giúp máy tính hiểu ngữ nghĩa và mối quan hệ giữa các từ trong ngôn ngữ tự nhiên. MPNet được thiết kế để đối mặt với thách thức biểu diễn văn bản đa ngôn ngữ, từ đó hỗ trợ nhiều ứng dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên.

c. Mô hình RAG với Vinallama



1. Chuẩn Bị Dữ Liệu

Dữ liệu từ các văn bản (document) được tải vào cơ sở dữ liệu vector. Điều này đảm bảo rằng dữ liệu có sẵn và có thể được truy xuất một cách nhanh chóng.

2. Quá Trình Truy Vấn

Người dùng đưa ra query, câu query đó sau khi được embedding sẽ đi vào index, hệ thống sẽ tìm kiếm các vector khác trong không gian vector đã được chỉ mục trước đó mà gần với vector của câu truy vấn và chuyển đổi thành dạng text.

3. Giao Tiếp với LLM

Các từ sau được truy vấn sẽ kết hợp cùng prompt vào mô hình Vinallama. Tại đây mô hình sẽ kết hợp các từ được truy xuất và phản hồi của chính nó đối với query thành câu trả lời cuối cùng.

2.2. Hướng tiếp cận Llama2 + NLLB (translation model)

a. Mô hình Llama2

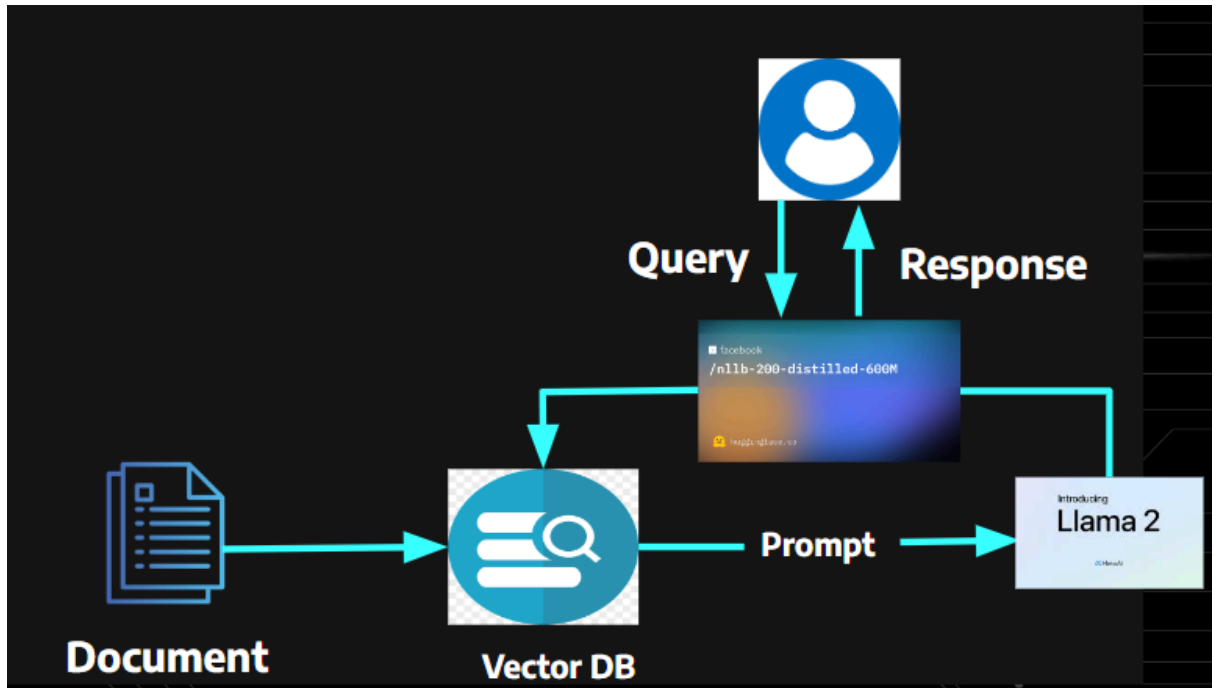
LLaMa-2 được giới thiệu như một mô hình ngôn ngữ AI mã nguồn mở được phát triển và giới thiệu bởi Meta Platforms và Microsoft. Đây là phiên bản tiếp theo của mô hình LLaMa, được tạo ra bởi Facebook AI Research và đội ngũ kỹ sư của họ. Kiến trúc của LLaMa2 là các Generative Pretrained Transformer dựa trên kiến trúc Transformer. LLaMa-2 mang lại nhiều cải tiến quan trọng so với phiên bản trước đó để nâng cao hiệu suất và khả năng của nó trong xử lý ngôn ngữ tự nhiên. Llama 2 được đào tạo với lượng dữ liệu nhiều hơn 40% so với phiên bản tiền nhiệm và có độ dài ngữ cảnh gấp đôi (4k). Llama2 gồm 4 mô hình khác nhau về số lượng tham số: 7B, 13B, 34B, và 70B. Ngoài ra có phiên bản Chatbot của Llama2 chính là Llama-2-Chat. Đối với mô hình Llama-2-Chat, Meta đã phát minh ra Ghost Attention (GAttn), một phương pháp tinh chỉnh mới.

b. Mô hình NLLB

Mô hình dịch ngôn ngữ NLLB (No Language Left Behind) là một mô hình trí tuệ nhân tạo được phát triển bởi Meta (Facebook). Được tạo ra để cải thiện chất lượng bản dịch, NLLB sử dụng các phương pháp học máy và mô hình ngôn ngữ neural network. NLLB có khả năng xử lý nhiều ngôn ngữ, giúp cải thiện hiệu suất trong môi trường đa ngôn ngữ, đặc biệt là trong ngữ cảnh của metaverse. Mô hình này có khả năng xây dựng các công nghệ hoạt động tốt trên nhiều loại ngôn ngữ, với hiệu suất cao và khả năng chuyển đổi chất lượng bản dịch lên đến 44%. NLLB bao gồm hơn 200 ngôn ngữ, không chỉ tập trung vào các ngôn ngữ phổ biến mà còn bao gồm các ngôn ngữ hiếm của châu Phi và châu Úc. Mục tiêu cuối cùng là cung cấp một phương tiện giao tiếp đa ngôn ngữ toàn diện.

Sử dụng kết hợp các phương pháp và mô hình ngôn ngữ neural network, NLLB không chỉ tập trung vào việc chuyển đổi giữa ngôn ngữ mà còn đảm bảo chất lượng và sự đa dạng trong việc phục vụ cộng đồng ngôn ngữ đa dạng.

c. Mô hình RAG với Llama2



1. Chuẩn Bị Dữ Liệu

Dữ liệu từ các văn bản (document) được tải vào index.

2. Quá Trình Truy Vấn

Khi user đưa ra câu query bằng tiếng việt, câu query sẽ qua mô hình translate NLLB để chuyển sang tiếng Anh.

3. Giao Tiếp với LLM

Sau khi đi qua mô hình Llama2, câu response sẽ được truyền vào mô hình NLLB để dịch từ tiếng Anh qua tiếng Việt, sau đó trả kết quả về cho người dùng.

3. So sánh hai hướng tiếp cận

Nhóm xây dựng RAG cho bộ luật đất đai, dữ liệu dạng file pdf (gồm hai bản tiếng Việt và tiếng Anh).

Link data: [data](#)

Với mỗi câu hỏi, nhóm sẽ dựa trên câu trả lời từ hai mô hình để đánh giá hai hướng tiếp cận này với nhau. Nhóm tiến hành trên bốn test case như bên dưới.

3.1. Test case 1

RAG Luật đất đai

Input

Đất thuộc quyền sở hữu của ai?

Clear

Submit

Output VinaLlama

Đất thuộc quyền sở hữu của người sử dụng đất (gồm người sử dụng đất khác với người sử dụng đất là hộ gia đình).

page_label: 13
file_path: data/vietnamese/luat_dat_dai.pdf

Điều 42. Quyền của người sử dụng đất
1. Người sử dụng đất được hưởng đầy đủ các quyền sau đây:
a) Sử dụng đất theo mục đích sử dụng đất đã được ghi trong Giấy chứng nhận quyền sử dụng đất hoặc trong giấy tờ về quyền sử dụng đất khác;
b) Được tiếp tục sử dụng đất theo mục đích sử dụng đất đã được ghi trong Giấy chứng nhận quyền sử dụng đất hoặc trong giấy tờ về quyền sử dụng đất khác nếu có nhu cầu;
c) Sử dụng đất để sản xuất nông nghiệp, lâm nghiệp, nuôi trồng thủy sản, làm muối;
d) Được chuyển nhượng quyền sử dụng đất cho người sử dụng đất khác;
đ) Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
e) Được tặng cho đất cho người sử dụng đất khác;
g) Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
h) Được tặng cho đất cho người sử dụng đất khác;
i) Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
k) Được tặng cho đất cho người sử dụng đất khác;
l) Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;

Output Llama2

Dựa trên văn bản được đưa ra, chúng ta không thể xác định được liệu đất thuộc về Ai hay không. Văn bản chỉ cung cấp thông tin về chế độ sử dụng đất và các loại đất khác nhau, bao gồm các vùng công nghệ cao.

Câu hỏi (Input): Đất thuộc quyền sở hữu của ai?

Output VinaLlama:

Đất thuộc quyền sở hữu của người sử dụng đất (gồm người sử dụng đất khác với người sử dụng đất là hộ gia đình).

page_label: 13

file_path: data/vietnamese/luat_dat_dai.pdf

Điều 42. Quyền của người sử dụng đất

1. Người sử dụng đất được hưởng đầy đủ các quyền sau đây:

- Sử dụng đất theo mục đích sử dụng đất đã được ghi trong Giấy chứng nhận quyền sử dụng đất hoặc trong giấy tờ về quyền sử dụng đất khác;
- Được tiếp tục sử dụng đất theo mục đích sử dụng đất đã được ghi trong Giấy chứng nhận quyền sử dụng đất hoặc trong giấy tờ về quyền sử dụng đất khác nếu có nhu cầu;
- Sử dụng đất để sản xuất nông nghiệp, lâm nghiệp, nuôi trồng thủy sản, làm muối;
- Được chuyển nhượng quyền sử dụng đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- Được tặng cho đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- Được tặng cho đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- Được tặng cho đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- Được tặng cho đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- Được tặng cho đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- Được tặng cho đất cho người sử dụng đất khác;
- Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;

- u) Được tặng cho đất cho người sử dụng đất khác;
- v) Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- w) Được tặng cho đất cho người sử dụng đất khác;
- x) Được nhận thừa kế phần đất thừa kế khi người sử dụng đất chết;
- y) Được tặng cho đất cho người sử dụng

Output Llama2:

Dựa trên văn bản được đưa ra, chúng ta không thể xác định được liệu đất thuộc về Ai hay không. văn bản chỉ cung cấp thông tin về chế độ sử dụng đất và các loại đất khác nhau, bao gồm các vùng công nghệ cao.

3.2. Test case 2

RAG Luật đất đai

VinaLlama
Llama2

Input

Đối tượng áp dụng của luật đất đai?

Output

Đối tượng áp dụng của luật đất đai là các chủ thể sau đây:

1. Tổ chức, cá nhân, người Việt Nam định cư ở nước ngoài, doanh nghiệp có vốn đầu tư nước ngoài được phép thực hiện dự án thăm dò, khai thác khoáng sản;
2. Tổ chức kinh tế có một phần đất sử dụng cho hoạt động khoáng sản theo quy định của pháp luật về đất đai;
3. Doanh nghiệp nhà nước, tổ chức chính trị - xã hội, tổ chức kinh tế do Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất hoặc tài sản gắn liền với đất để thực hiện dự án đầu tư;
4. Cơ quan quản lý nhà nước về đất đai;
5. Tổ chức nước ngoài được Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất để thực hiện dự án đầu tư theo quy định của pháp luật về đầu tư;
6. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
7. Người Việt Nam định cư ở nước ngoài được Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất để thực hiện dự án đầu tư;
8. Tổ chức kinh tế là doanh nghiệp nhà nước, tổ chức chính trị - xã hội, tổ chức kinh tế do Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất hoặc tài sản gắn liền với đất để thực hiện dự án đầu tư.
9. Người Việt Nam định cư ở nước ngoài được Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất để thực hiện dự án đầu tư;
10. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
11. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
12. Tổ chức kinh tế là doanh nghiệp nhà nước, tổ chức chính trị - xã hội, tổ chức kinh tế do Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất hoặc tài sản gắn liền với đất để thực hiện dự án đầu tư.
13. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
14. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
15. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;

Run

RAG Luật đất đai

VinaLlama
Llama2

Input

Đối tượng áp dụng của luật đất đai?

Output

Dựa trên đoạn trích được cung cấp từ luật đất đai Việt Nam, đối tượng của luật đất đai áp dụng dường như là Nhà nước và đại diện của nó, cũng như các cá nhân và tổ chức sử dụng hoặc muốn sử dụng đất ở Việt Nam.

Run

Câu hỏi (Input): Đối tượng áp dụng của luật đất đai?

Output VinaLlama:

Đối tượng áp dụng của luật đất đai là các chủ thể sau đây:

1. Tổ chức, cá nhân, người Việt Nam định cư ở nước ngoài, doanh nghiệp có vốn đầu tư nước ngoài được phép thực hiện dự án thăm dò, khai thác khoáng sản;
2. Tổ chức kinh tế có một phần đất sử dụng cho hoạt động khoáng sản theo quy định của pháp luật về đất đai;
3. Doanh nghiệp nhà nước, tổ chức chính trị - xã hội, tổ chức kinh tế do Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất hoặc tài sản gắn liền với đất để thực hiện dự án đầu tư;
4. Cơ quan quản lý nhà nước về đất đai;

5. Tổ chức nước ngoài được Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất để thực hiện dự án đầu tư theo quy định của pháp luật về đầu tư;
6. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
7. Người Việt Nam định cư ở nước ngoài được Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất để thực hiện dự án đầu tư;
8. Tổ chức kinh tế là doanh nghiệp nhà nước, tổ chức chính trị - xã hội, tổ chức kinh tế do Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất hoặc tài sản gắn liền với đất để thực hiện dự án đầu tư.
9. Người Việt Nam định cư ở nước ngoài được Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất để thực hiện dự án đầu tư;
10. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
11. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
12. Tổ chức kinh tế là doanh nghiệp nhà nước, tổ chức chính trị - xã hội, tổ chức kinh tế do Nhà nước giao đất, cho thuê đất, chuyển nhượng quyền sử dụng đất hoặc tài sản gắn liền với đất để thực hiện dự án đầu tư.
13. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
14. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;
15. Tổ chức nước ngoài được Nhà nước cho thuê đất, cho thuê lại đất để thực hiện dự án đầu tư;

Output Llama2:

Dựa trên đoạn trích được cung cấp từ luật đất Việt Nam, đối tượng của luật đất áp dụng dường như là Nhà nước và đại diện của nó, cũng như các cá nhân và tổ chức sử dụng hoặc muốn sử dụng đất ở Việt Nam.

3.3. Test case 3

RAG Luật đất đai

Input

Luật nào quan trọng nhất trong luật đất đai?

Clear

Submit

Output VinaLlama

(Điều 182)
1. Đất quốc phòng, an ninh và đất phục vụ quốc phòng, an ninh bao gồm đất do Nhà nước quản lý trực tiếp, đất do Nhà nước giao, cho thuê, cho mượn hoặc được Nhà nước hỗ trợ kinh phí để quản lý, bảo vệ.
2. Đất quốc phòng, an ninh bao gồm:
a) Đất phục vụ quốc phòng, an ninh;
b) Đất có diện tích từ 500 ha trở lên thuộc khu vực I, khu vực II, khu vực III, khu vực IV;
c) Đất có diện tích từ 1.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
d) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
đ) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
e) Đất có diện tích từ 10.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
g) Đất có diện tích từ 3.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
h) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
i) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
k) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
l) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
m) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
n) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
o) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
p) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
q) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;
r) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

Output Llama2

Luật quan trọng nhất của đất là Luật đất của Việt Nam, luật này quy định quyền và nghĩa vụ của người sử dụng đất, bao gồm cả nhân, tổ chức và nhà nước, và cung cấp một khuôn khổ pháp lý cho quản lý và sử dụng đất ở Việt Nam. Nó bao gồm nhiều khía cạnh khác nhau của quản lý đất, chẳng hạn như phân bổ đất, thuê đất, lấy lại đất, và giá đất, và nêu rõ trách nhiệm của nhà nước và chính quyền địa phương trong việc quản lý và giám sát việc sử dụng đất.

Flag

Câu hỏi (Input): Luật nào quan trọng nhất trong luật đất đai?

Output VinaLlama:

Luật Đất đai 2013

9. Đất quốc phòng, an ninh và đất phục vụ quốc phòng, an ninh
(Điều 182)

1. Đất quốc phòng, an ninh và đất phục vụ quốc phòng, an ninh bao gồm đất do Nhà nước quản lý trực tiếp, đất do

Nhà nước giao, cho thuê, cho mượn hoặc được Nhà nước hỗ trợ kinh phí để quản lý, bảo vệ.

2. Đất quốc phòng, an ninh bao gồm:

a) Đất phục vụ quốc phòng, an ninh;

b) Đất có diện tích từ 500 ha trở lên thuộc khu vực I, khu vực II, khu vực III, khu vực IV;

c) Đất có diện tích từ 1.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

d) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

đ) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

e) Đất có diện tích từ 10.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

g) Đất có diện tích từ 3.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

h) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

i) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

k) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

l) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

m) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

n) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

o) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

p) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

q) Đất có diện tích từ 5.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

r) Đất có diện tích từ 2.000 ha trở lên thuộc khu vực I, khu vực II, khu vực III;

Output Llama2:

Luật quan trọng nhất của đất là Luật đất của Việt Nam, luật này quy định quyền và nghĩa vụ của người sử dụng đất, bao gồm cá nhân, tổ chức và nhà nước, và cung cấp một khuôn khổ pháp lý cho quản lý và sử dụng đất ở Việt Nam. Nó bao gồm nhiều khía cạnh khác nhau của quản lý đất, chẳng hạn như phân bổ đất, thuê đất, lấy lại đất, và giá đất, và nêu rõ trách nhiệm của nhà nước và chính quyền địa phương trong việc quản lý và giám sát việc sử dụng đất.

3.4. Test case 4

RAG Luật đất đai

VinaLlama

Llama2

Input

Bồi thường khi nhà nước thu hồi đất?

Output

- Nhà nước thu hồi đất để phát triển kinh tế - xã hội vì lợi ích quốc gia, công cộng theo Điều 60 Luật Đất đai năm 2013.

- Khoản 1 Điều 66. Nghị định số 43/2014/NĐ-CP ngày 15/5/2014 quy định về thu hồi đất để thực hiện các dự án quan trọng quốc gia do Quốc hội quyết định chủ trương đầu tư mà phải thu hồi đất.

- Khoản 2 Điều 66. Nghị định số 43/2014/NĐ-CP quy định về thu hồi đất để thực hiện các dự án do Thủ tướng Chính phủ chấp thuận, quyết định đầu tư mà phải thu hồi đất, bao gồm:

a) Dự án xây dựng khu công nghiệp, khu chế xuất, khu công nghệ cao, khu kinh tế; khu đô thị mới, dự án đầu tư bằng nguồn vốn hỗ trợ phát triển chính thức (ODA).

b) Dự án xây dựng trụ sở cơ quan nhà nước, tổ chức chính trị, tổ chức chính trị - xã hội ở trung ương; trụ sở của tổ chức nước ngoài có chức năng ngoại giao; công trình di tích lịch sử - văn hóa, danh lam thắng cảnh được xếp hạng, công viên, quảng trường, tượng đài, bia tưởng niệm, công trình sự nghiệp công cấp quốc gia;

c) Dự án xây dựng kết cấu hạ tầng kỹ thuật cấp quốc gia gồm giao thông, thủy lợi, cấp nước, thoát nước, điện lực, thông tin liên lạc; hệ thống dẫn, chứa xăng dầu, khí đốt; kho dự trữ quốc gia; công trình thu gom, xử lý chất thải;

d) Dự án xây dựng kết cấu hạ tầng kỹ thuật cấp tỉnh gồm giao thông, thủy lợi, cấp nước, thoát nước, điện lực, thông tin liên lạc; hệ thống dẫn, chứa xăng dầu, khí đốt; kho dự trữ quốc gia; công trình thu gom, xử lý chất thải.

Run

RAG Luật đất đai

VinaLlama

Llama2

Input

Bồi thường khi nhà nước thu hồi đất?

Output

Theo Luật Đất của Việt Nam, khi nhà nước đòi lại đất, người sử dụng đất có quyền được bồi thường, bồi thường phải được thực hiện dưới dạng phân bổ đất mới với cùng mục đích sử dụng đất với đất được phục hồi. Nếu không có đất có sẵn để bồi thường, người sử dụng đất sẽ nhận được bồi thường bằng tiền bạc được tính theo giá đất cụ thể của loại đất được phục hồi mà Ủy ban Nhân dân tỉnh quyết định vào thời điểm quyết định phục hồi. Bồi thường phải được thực hiện theo cách dân chủ, không thiên vị, bình đẳng, công khai, kịp thời và hợp pháp.

Run

Câu hỏi (Input): Bồi thường khi nhà nước thu hồi đất?

Output VinaLlama:

- Nhà nước thu hồi đất để phát triển kinh tế - xã hội vì lợi ích quốc gia, công cộng theo Điều 60 Luật Đất đai năm 2013.

- Khoản 1 Điều 66. Nghị định số 43/2014/NĐ-CP ngày 15/5/2014 quy định về thu hồi đất để thực hiện các dự án quan trọng quốc gia do Quốc hội quyết định chủ trương đầu tư mà phải thu hồi đất.

- Khoản 2 Điều 66. Nghị định số 43/2014/NĐ-CP quy định về thu hồi đất để thực hiện các dự án do Thủ tướng Chính phủ chấp thuận, quyết định đầu tư mà phải thu hồi đất, bao gồm:

a) Dự án xây dựng khu công nghiệp, khu chế xuất, khu công nghệ cao, khu kinh tế; khu đô thị mới, dự án đầu tư bằng nguồn vốn hỗ trợ phát triển chính thức (ODA).

b) Dự án xây dựng trụ sở cơ quan nhà nước, tổ chức chính trị, tổ chức chính trị - xã hội ở trung ương; trụ sở của tổ chức nước ngoài có chức năng ngoại giao; công trình di tích lịch sử - văn hóa, danh lam thắng cảnh được xếp hạng, công viên, quảng trường, tượng đài, bia tưởng niệm, công trình sự nghiệp công cấp quốc gia;

c) Dự án xây dựng kết cấu hạ tầng kỹ thuật cấp quốc gia gồm giao thông, thủy lợi, cấp nước, thoát nước, điện lực, thông tin liên lạc; hệ thống dẫn, chứa xăng dầu, khí đốt; kho dự trữ quốc gia

gia; công trình thu gom, xử lý chất thải;

d) Dự án xây dựng kết cấu hạ tầng kỹ thuật cấp tỉnh gồm giao thông, thủy lợi, cấp nước, thoát nước, điện lực, thông tin liên lạc; hệ thống dẫn, chứa xăng dầu, khí đốt; kho dự trữ quốc gia; công trình thu gom, xử lý chất thải.

Output Llama2:

Theo Luật Đất của Việt Nam, khi nhà nước đòi lại đất, người sử dụng đất có quyền được bồi thường. bồi thường phải được thực hiện dưới dạng phân bổ đất mới với cùng mục đích sử dụng đất với đất được phục hồi. nếu không có đất có sẵn để bồi thường, người sử dụng đất sẽ nhận được bồi thường bằng tiền bạc được tính theo giá đất cụ thể của loại đất được phục hồi mà Ủy ban Nhân dân tỉnh quyết định vào thời điểm quyết định phục hồi. bồi thường phải được thực hiện theo cách dân chủ, không thiên vị, bình đẳng, công khai, kịp thời và hợp pháp.

3.5. Nhận xét

- Với hướng tiếp cận ViniLlama + mpnet, mô hình trả về kết quả dài dòng, lan man, các ý bị lặp lại, nhưng chỉ có vài ý đúng.
- Với hướng tiếp cận Llama2 + NLLB, mô hình trả về kết quả ngắn gọn, được tóm tắt sẵn nội dung.

4. Demo

Nhóm sử dụng Gradio để demo hai mô hình của hai hướng tiếp cận.

Link Colab:  [app.ipynb](#)

4.1. Gradio

Gradio là một thư viện mã nguồn mở trong Python giúp xây dựng giao diện người dùng cho mô hình máy học một cách nhanh chóng. Gradio thường kết hợp mạnh mẽ với Hugging Face, một hệ sinh thái cung cấp các mô hình ngôn ngữ tự nhiên đa dạng. Có thể tích hợp mô hình RAG từ Hugging Face vào Gradio để tạo ra ứng dụng có khả năng tìm kiếm thông tin và tạo ra câu trả lời tự động.

Các Bước Xây Dựng Ứng Dụng RAG với Gradio và Hugging Face:

1. **Chuẩn Bị Môi Trường:** Bắt đầu bằng cách cài đặt Gradio và Hugging Face trên Google Colab.
2. **Chọn và Tải Mô Hình RAG từ Hugging Face:** Chọn một mô hình RAG từ Hugging Face Hub, sau đó sử dụng API để tải mô hình về. Đảm bảo chọn mô hình tương thích với nhu cầu ứng dụng của bạn.
3. **Xây Dựng Giao Diện Người Dùng với Gradio:** Sử dụng Gradio để tạo giao diện người dùng tương tác cho ứng dụng RAG của bạn. Định nghĩa các hàm xử lý và tích hợp chúng vào giao diện.
4. **Chạy Ứng Dụng trên Google Colab:** Chạy ứng dụng và tương tác với mô hình RAG thông qua giao diện Gradio trên Google Colab.

```
# DEMO1
def get_response(text):
    response1 = query_engine.query(text)
    response2 = query(text)
    return [response1, response2]

# Gradio
interface = gr.Interface(get_response, inputs=gr.Text(label="Input"), outputs=[gr.Text(label="Output VinaLlama"), gr.Text(label="Output Llama2")], title = 'RAG Luật đất đai')
interface.launch(share=True)
```

```
# DEMO2
with gr.Blocks() as demo:
    gr.Markdown('RAG Luật đất đai')
    with gr.Tab("VinaLlama"):
        text_input1 = gr.Textbox(label="Input")
        text_output1 = gr.Textbox(label="Output")
        run1 = gr.Button("Run")
    with gr.Tab("Llama2"):
        text_input2 = gr.Textbox(label="Input")
        text_output2 = gr.Textbox(label="Output")
        run2 = gr.Button("Run")

    run1.click(query_engine.query, inputs=text_input1, outputs=text_output1)
    run2.click(query, inputs=text_input2, outputs=text_output2)

demo.launch(share=True)
```

4.2. Giao diện

Nhóm tiến hành thực hiện trên hai bản demo.

4.2.1. DEMO 1

Với mỗi Input, sẽ tiến hành trả về cùng lúc hai Ouput của hai mô hình.

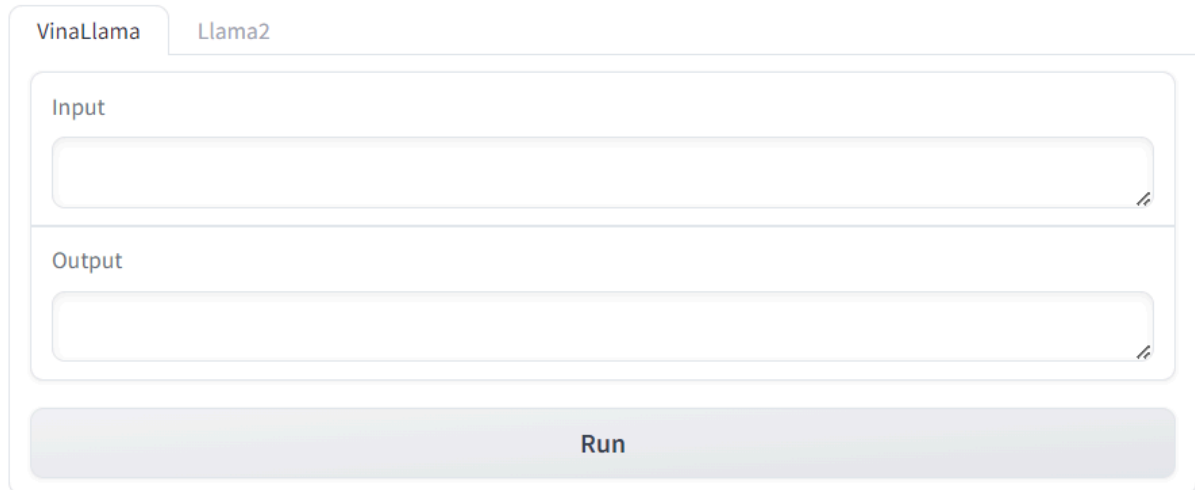
RAG Luật đất đai

The image shows a web interface for a RAG system titled "RAG Luật đất đai". The interface is split into two columns. The left column has an "Input" text area, a "Clear" button, and a "Submit" button. The right column has two output text areas labeled "Output VinaLlama" and "Output Llama2", and a "Flag" button.

4.2.2. DEMO 2

Có thể lựa chọn mô hình muốn thực hiện, sau đó với Input, sẽ trả về Output của mô hình đã được lựa chọn.

RAG Luật đất đai



The interface shows a selection of models at the top: 'VinaLlama' and 'Llama2'. Below this, there are two text input areas labeled 'Input' and 'Output'. A 'Run' button is positioned at the bottom of the input section.

4.3. Cách sử dụng

Bước 1: Vào url tương ứng với mỗi demo, url do Gradio tạo sẽ tồn tại trong 72 giờ.

Bước 2:

- Với DEMO 1: Nhập câu hỏi vào ô Input và bấm Run. Kết quả sẽ được hiển thị ở hai ô Output.
- Với DEMO 2: Lựa chọn mô hình muốn thực hiện (VinaLlama hoặc Llama2). Sau đó nhập câu hỏi vào ô Input và bấm Run. Kết quả sẽ được hiển thị ở hai ô Output. Trong quá trình đợi kết quả của mô hình này, có thể bấm lựa chọn mô hình còn lại và tiến hành tương tự, nhập Input và bấm Run, đợi dự đoán.

REFERENCES

- [1] High-Level Concepts - LLaMaIndex v0.10.11.Post1. (n.d.).
https://docs.llamaindex.ai/en/stable/getting_started/concepts.html
- [2] Dash, S. K. (2024, February 7). *Build a RAG pipeline with the LLAMA Index*. Analytics Vidhya.
<https://www.analyticsvidhya.com/blog/2023/10/rag-pipeline-with-the-llama-index/>
- [3] Nhẽo C. T. N. (2023, July 24). Llama 2 là gì? Sử dụng Llama 2 như thế nào? Quantrimang.com. <https://quantrimang.com/lang-cong-nghe/llama-2-la-gi-198309>
- [4] AI breakdown or: Takeaways from the 78-page Llama-2 paper | Deepgram. (n.d.). Deepgram.
<https://deepgram.com/learn/llama-2-paper-explained>
- [5] BKAI-HUST Foundation Models Lab. Llama-2-bk, 2023.
- [6] Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning, 2023.
- [7] NLLB. (n.d.). https://huggingface.co/docs/transformers/model_doc/nllb
- [8] Quan Nguyen, Huy Pham and Dung Dao. VinaLLaMA: LLaMA-based Vietnamese Foundation Model. arXiv:2312.11011.
- [9] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In NIPS, pages 5754–5764, 2019.
- [10] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. arXiv:2004.09297.