MỤC LỤC

Table of Contents

LỜI CẢM ƠN

Trước hết, tôi xin chân thành cảm ơn đến các thầy cô tại Viện Trí tuệ Nhân tạo, Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội. Nhờ sự tận tâm trong giảng dạy và hướng dẫn, cùng với những kiến thức chuyên môn và kỹ năng thực tiễn mà các thầy cô đã truyền đạt, tôi đã có cơ hội lĩnh hội nhiều kiến thức quý báu. Những kiến thức đó không chỉ giúp tôi phát triển tư duy mà còn tạo nền tảng vững chắc cho con đường học tập và nghề nghiệp của mình.

Tôi xin đặc biệt cảm ơn đến thầy **TS. Trần Văn Khánh**, trong suốt học kỳ vừa rồi thầy đã đồng hành hướng dẫn tôi trong quá trình làm dự án này. Tôi biết thêm được nhiều kiến thức mới đặc biệt là trong lĩnh vực Xử lý ngôn ngữ tự nhiên, có được những góc nhìn mới, toàn diện hơn để chuẩn bị hành trang cho con đường nghiên cứu sau này.

I. Giới thiệu chung và yêu cầu bài toán

Bài toán: Hỏi đáp về luật tiếng việt

Mục tiêu: Xây dựng một hệ thống thông minh hỗ trợ người dùng trả lời các câu hỏi liên quan đến pháp luật, cung cấp thông tin chính xác, cập nhật và dễ hiểu dựa trên các tài liệu chính thức, quy định pháp lý và thông tin thời sự.

Pháp luật là một lĩnh vực đòi hỏi tính chính xác cao và thông tin phải luôn được cập nhật để phù hợp với các thay đổi liên tục. Người dùng thường gặp khó khăn trong việc tiếp cận các văn bản pháp lý dài và phức tạp hoặc không có đủ chuyên môn để hiểu rõ các điều khoản. Do đó, hệ thống hỏi đáp pháp luật cần được thiết kế để giải quyết vấn đề này, hỗ trợ người dùng tra cứu nhanh chóng, hiệu quả.

Với sự phát triển của xử lý ngôn ngữ tự nhiên (NLP) đặc biệt là các mô hình ngôn ngữ lớn (LLM), chúng đã trở thành phương hướng tiếp cận chủ yếu cho bài toán này nhờ vào khả năng học hỏi, hiểu ngôn ngữ thông qua việc đào tạo trên một tập dữ liệu lớn, đa dạng.

Mặc dù LLM (Large Language Models) đã chứng tỏ được khả năng vượt trội trong xử lý ngôn ngữ tự nhiên, nhưng khi áp dụng vào bài toán pháp luật, chúng vẫn tồn tại các hạn chế sau:

1. **Thiếu khả năng cập nhật thông tin**:
   Các mô hình LLM thường được huấn luyện trên dữ liệu cố định và không thể tự động cập nhật các thay đổi mới trong hệ thống pháp luật. Điều này dẫn đến nguy cơ cung cấp thông tin lỗi thời, đặc biệt là trong các tình huống pháp lý quan trọng.

2. **Thông tin không chính xác hoặc thiếu căn cứ (vấn đề "ảo giác")**:
   LLM đôi khi có thể tạo ra nội dung có vẻ hợp lý nhưng thực chất không chính xác, gây hiểu nhầm cho người dùng. Trong lĩnh vực pháp luật, sự mơ hồ hoặc sai sót thông tin có thể dẫn đến hậu quả nghiêm trọng.

3. **Không có khả năng truy xuất nguồn gốc**:
   LLM không cung cấp nguồn tham chiếu cho các câu trả lời. Điều này làm giảm tính minh bạch và độ tin cậy của hệ thống trong các tình huống đòi hỏi bằng chứng pháp lý hoặc căn cứ rõ ràng.

4. **Hạn chế xử lý các văn bản phức tạp**:
   Các tài liệu pháp luật thường dài và đòi hỏi khả năng hiểu sâu sắc ngữ cảnh, điều mà LLM đôi khi không thể xử lý chính xác.

Giải pháp đề xuất: Tích hợp RAG và Search Web

Để khắc phục các hạn chế của LLM, dự án đề xuất tích hợp phương pháp **RAG (Retrieval-Augmented Generation)** và **Search Web** nhằm tăng cường hiệu quả hệ thống.

1. **Tăng cường tính chính xác**:
   RAG cho phép hệ thống kết hợp khả năng xử lý ngôn ngữ tự nhiên của LLM với các tài liệu pháp luật được truy xuất từ cơ sở dữ liệu hoặc web. Câu trả lời được xây dựng dựa trên các tài liệu pháp lý cụ thể, giúp đảm bảo tính chính xác.

2. **Đảm bảo tính cập nhật**:
   Việc tích hợp Search Web giúp hệ thống truy cập thông tin mới nhất, bao gồm các thay đổi, sửa đổi hoặc bổ sung trong luật pháp. Điều này giúp duy trì tính thời sự và phù hợp của thông tin cung cấp.

3. **Truy xuất nguồn gốc**:
   Hệ thống có thể cung cấp câu trả lời kèm theo nguồn tài liệu hoặc đường dẫn tham khảo, tăng độ tin cậy và cho phép người dùng tự kiểm chứng thông tin.

4. **Xử lý các câu hỏi phức tạp**:
   RAG khai thác hiệu quả từ các tài liệu dài hoặc phức tạp, đảm bảo các câu trả lời phản ánh đầy đủ và chính xác nội dung pháp lý liên quan.

II. Tóm tắt lý thuyết, giải pháp

Các lý thuyết, giải pháp, thuật toán liên quan

1. LLMs

   Large Language Models (LLMs), như GPT hoặc các biến thể tương tự, là những mô hình AI được huấn luyện trên lượng lớn dữ liệu văn bản. Trong bối cảnh bài toán hỏi đáp luật pháp, LLM có thể đóng vai trò như một công cụ mạnh mẽ hỗ trợ tra cứu, giải thích và cung cấp câu trả lời cho các câu hỏi liên quan đến luật pháp, quy định và các văn bản pháp lý.

2. Fine-tuning

   Fine-tuning là quá trình tiếp tục huấn luyện một mô hình ngôn ngữ lớn (LLM) đã được tiền huấn luyện (pre-trained) trên một tập dữ liệu cụ thể, nhằm làm cho mô hình hoạt động tốt hơn trong một lĩnh vực hoặc ngữ cảnh nhất định. Trong bối cảnh hỏi đáp luật pháp, fine-tuning trên dữ liệu pháp lý cụ thể là một cách hiệu quả để cải thiện độ chính xác và tính hữu ích của mô hình.

   Tuy nhiên, có thể thấy phương pháp này đòi hỏi tập dữ liệu lớn và chất lượng, tốn nhiều thời gian.

3. RAG

   RAG (Retrieval-Augmented Generation) là phương pháp kết hợp truy xuất thông tin từ cơ sở dữ liệu với khả năng sinh văn bản của mô hình ngôn ngữ lớn (LLM) nhằm cải thiện độ chính xác và tính thực tế trong câu trả lời.

   Các tài liệu (documents) sau khi được thu thập, sẽ được chunking (khoảng 500-1200 tokens, có overlap), sau đó sẽ được đi qua model embedding để chuyển thành semantic vector và lưu trong cơ sở dữ liệu (ví dụ như ChromaDB, NanoDB,...). Khi nhận được query, nó cũng sẽ được đi qua cùng model embedding và sử dụng Cosine Search (ANN) để tìm ra các vector gần nhất.

   Tuy đơn giản, phương pháp này cũng có một số hạn chế. Nó phụ thuộc vào “flat data representation”, nghĩa là các chunk chỉ là những dữ liệu

thô, không có liên kết với nhau, vì vậy khi gặp câu hỏi khó, tổng quát, nó chỉ có thể truy xuất được những thông tin riêng lẻ, không quá liên kết với nhau.

4. Search Tools

   Search Tool trong bối cảnh truy xuất thông tin trên web là cách sử dụng các công cụ hoặc phương pháp để lấy thông tin từ các trang web hoặc cơ sở dữ liệu trực tuyến dựa trên câu hỏi hoặc yêu cầu đầu vào. Trong hệ thống RAG, search web thường được sử dụng để cập nhật thông tin thời gian thực từ Internet, bổ sung vào khả năng sinh câu trả lời của LLM.

5. Planning

   Planning là phương nhằm cải thiện khả năng của mô hình ngôn ngữ lớn (LLM) khi giải quyết các bài toán phức tạp, đòi hỏi suy luận nhiều bước hoặc hành động tương tác với môi trường. Các phương pháp này thường kết hợp suy luận logic, phản hồi ngữ cảnh, và khả năng tự điều chỉnh của mô hình. Một số phương pháp tiêu biểu bao gồm Chain of Thought (CoT), ReAct, và Reflexion.

   Chain of Thought là một kỹ thuật hướng dẫn mô hình suy luận theo từng bước, thay vì trả lời trực tiếp. Phương pháp này khuyến khích LLM viết ra quá trình suy nghĩ để giải quyết bài toán.

   ReAct kết hợp suy luận (Reasoning) và hành động (Acting ở đây là Seach Web) để giải quyết các nhiệm vụ đòi hỏi tương tác với môi trường hoặc sử dụng các công cụ.

   Reflexion tập trung vào khả năng tự đánh giá và cải thiện của mô hình, giúp nó học từ lỗi sai và điều chỉnh cách tiếp cận.

Giải pháp cho bài toán – GraphRAG, Search, Planning

1. GraphRAG (LightRAG)

LẬP CHỈ MỤC VĂN BẢN DỰA TRÊN ĐỒ THỊ

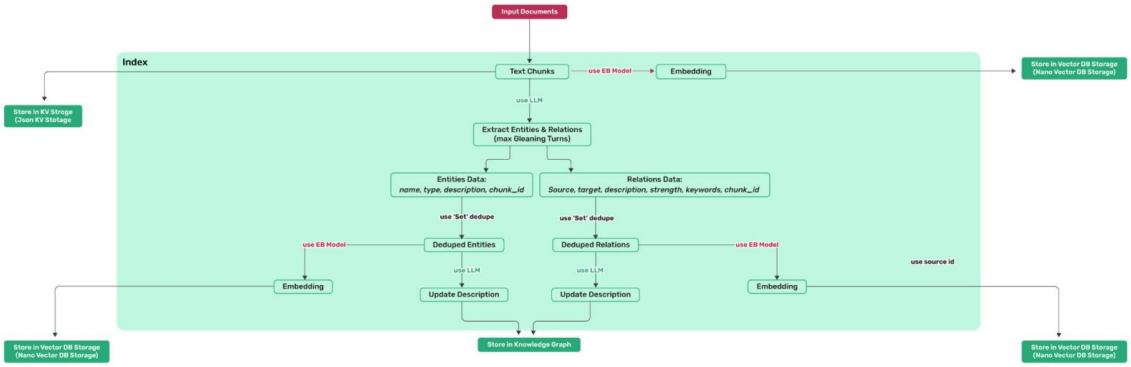
LightRAG sử dụng đồ thị để cải thiện việc trích xuất thực thể và mối quan hệ từ văn bản. Hệ thống chia nhỏ tài liệu thành các phần nhỏ để xử lý nhanh hơn, sau đó sử dụng các mô hình ngôn ngữ lớn (LLM) để nhận diện thực thể (như tên, ngày, địa điểm, sự kiện) và mối quan hệ giữa chúng.

Kết quả được chuyển thành đồ thị tri thức $D=(V,E)$, gồm:

- Trích xuất thực thể và quan hệ: Hàm $R(\cdot)$, yêu cầu LLM nhận diện các thực thể (nút) và quan hệ (cạnh). Ví dụ: trích xuất thực thể "Bác sĩ tim mạch" và "Bệnh tim" cùng quan hệ "Bác sĩ tim mạch chẩn đoán bệnh tim" từ văn bản.
- Tạo cặp khóa-giá trị: Hàm $P(\cdot)$, sử dụng LLM để tạo cặp khóa-giá trị (key-value) cho mỗi thực thể và quan hệ, giúp tối ưu hóa việc truy xuất.
- Loại bỏ dữ liệu trùng lặp: Hàm $D(\cdot)$ hợp nhất các thực thể và quan hệ trùng lặp, giảm kích thước đồ thị, cải thiện hiệu suất xử lý.

Lợi ích chính:

- Hiểu thông tin toàn diện: Đồ thị tri thức cho phép trích xuất thông tin liên kết qua nhiều phần tài liệu.
- Hiệu suất truy xuất vượt trội: Dữ liệu khóa-giá trị giúp truy xuất nhanh và chính xác hơn so với các phương pháp truyền thống.
- Cập nhật dữ liệu theo thời gian thực: LightRAG hỗ trợ cập nhật dữ liệu mới D' mà không cần tái xử lý toàn bộ. Dữ liệu mới được tích hợp bằng cách hợp nhất các nút V' và cạnh E' với đồ thị gốc, đảm bảo thông tin cập nhật mà không làm mất dữ liệu cũ.



MÔ HÌNH TRUY XUẤT HAI CẤP

LightRAG sử dụng chiến lược truy xuất hai cấp để xử lý câu hỏi cụ thể và khái quát:

- Truy vấn cụ thể: Tìm kiếm các thực thể hoặc mối quan hệ chi tiết, ví dụ: “Ai là tác giả của Pride and Prejudice?”
- Truy vấn khái quát: Xử lý các câu hỏi bao quát, như: “Trí tuệ nhân tạo ảnh hưởng đến giáo dục hiện đại như thế nào?”

LightRAG kết hợp đồ thị và biểu diễn vector để tối ưu hóa truy xuất:

- Trích xuất từ khóa: Xác định từ khóa cụ thể và khái quát từ truy vấn.
- So khớp từ khóa: Sử dụng cơ sở dữ liệu vector để so khớp từ khóa với thực thể và mối quan hệ.
- Tích hợp liên kết cao cấp: Truy xuất các nút và cạnh lân cận (liên kết một bước) để cung cấp thông tin toàn diện hơn.

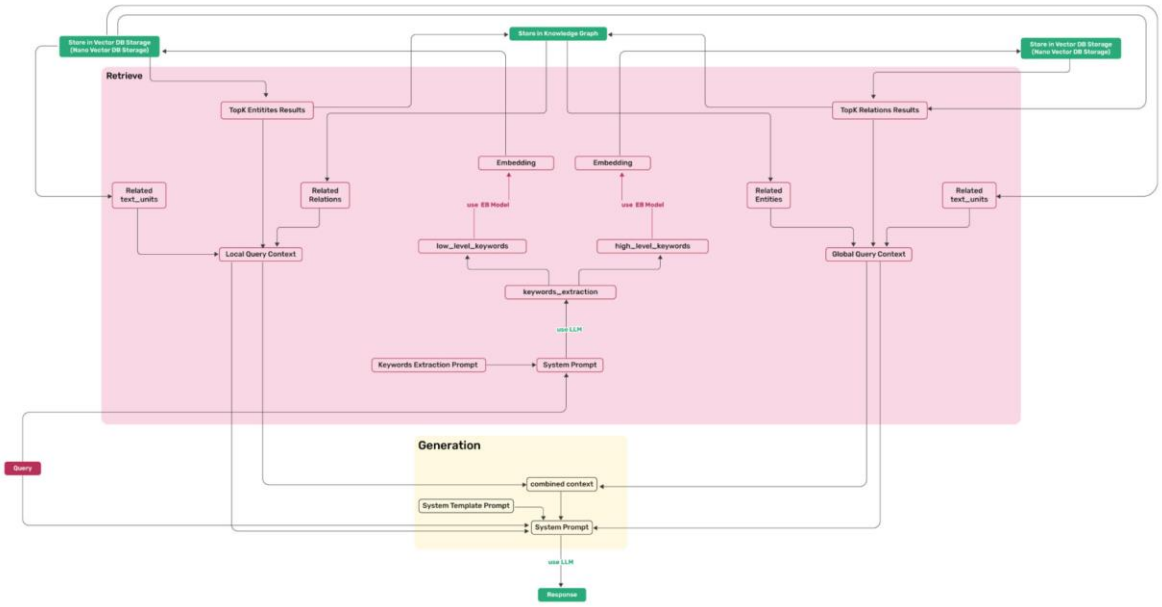
Lợi ích:

- Tích hợp thông tin từ đồ thị giúp trả lời chính xác và đầy đủ hơn.
- Kết hợp thông tin cục bộ và toàn cục giúp tối ưu hóa hiệu suất tìm kiếm.

TẠO CÂU TRẢ LỜI DỰA TRÊN TRUY XUẤT

LightRAG sử dụng dữ liệu truy xuất để tạo câu trả lời bằng cách:

- Tích hợp ngữ cảnh: LLM kết hợp các đoạn văn bản liên quan từ đồ thị (gồm tên, mô tả thực thể, mối quan hệ, trích đoạn gốc) để hiểu rõ ý định của truy vấn.
- Sinh câu trả lời: LLM tạo câu trả lời dựa trên thông tin tích hợp, đảm bảo phù hợp với nhu cầu người dùng.



Tóm lại, LightRAG là hệ thống RAG tiên tiến, kết hợp đồ thị tri thức và mô hình ngôn ngữ lớn để trích xuất, truy xuất, và trả lời thông tin hiệu quả, đồng thời hỗ trợ cập nhật dữ liệu mới một cách liền mạch.

2. Search

- Nhận câu hỏi từ người dùng: Người dùng nhập câu hỏi hoặc yêu cầu thông tin.
- Chuyển câu hỏi thành truy vấn tìm kiếm: Hệ thống chuyển đổi câu hỏi thành truy vấn tối ưu để gửi đến công cụ tìm kiếm (như Tavily Search, Duckduckgo, Google, Bing).

- Truy cập công cụ tìm kiếm: Sử dụng API công cụ tìm kiếm hoặc công cụ web scraping để gửi truy vấn và nhận kết quả.

- Phân tích kết quả tìm kiếm: Hệ thống nhận các kết quả (URLs, tiêu đề, đoạn tóm tắt). Chọn lọc kết quả liên quan nhất dựa trên mức độ phù hợp với truy vấn. Lấy nội dung từ các trang web được chọn.
- Tiền xử lý thông tin: Làm sạch nội dung web để loại bỏ các phần không cần thiết (như quảng cáo, menu). Trích xuất các đoạn thông tin hữu ích nhất để sử dụng làm ngữ cảnh.

3. Planning

   Sử dụng LLM có khả năng ngữ nghĩa mạnh để lên kế hoạch, điều phối, đánh giá hết quả. Nếu kết quả đánh giá không được tốt, chưa phù hợp với câu hỏi, thì chạy lại quả trình có sử dụng thêm search web.

III. Mô tả phần mềm cài đặt

NanoDB

NanoDB là một cơ sở dữ liệu nhẹ và tối ưu hóa, được thiết kế đặc biệt để lưu trữ và truy xuất các văn bản hoặc dữ liệu dạng text cho các ứng dụng AI và mô hình ngôn ngữ lớn (LLM). NanoDB cung cấp một giải pháp hiệu quả về tài nguyên và tốc độ, giúp lưu trữ lượng lớn dữ liệu văn bản mà không cần yêu cầu phần cứng hoặc phần mềm phức tạp.

Tavily Search

Tavily Search là một công cụ tìm kiếm được thiết kế để cung cấp trải nghiệm tìm kiếm nhanh chóng và hiệu quả, đặc biệt tập trung vào việc tìm kiếm thông tin từ các nguồn trực tuyến khác nhau, bao gồm cả web và các cơ sở dữ liệu lớn. Tavily Search có thể được sử dụng trong nhiều lĩnh vực, từ tìm kiếm văn bản đến truy xuất dữ liệu trong các ứng dụng AI.

Neo4j

Neo4j là một hệ quản trị cơ sở dữ liệu đồ thị (graph database), được thiết kế đặc biệt để lưu trữ và truy vấn dữ liệu dạng đồ thị, nơi các thực thể và mối quan hệ giữa chúng được biểu diễn rõ ràng. Với khả năng tối ưu hóa các truy vấn đồ thị, Neo4j đã trở thành một công cụ phổ biến trong nhiều ứng dụng, từ mạng xã hội, quản lý chuỗi cung ứng, đến các hệ thống đề xuất và phân tích dữ liệu phức tạp.

IV. Kết quả đạt được

Xử lý dữ liệu

Dữ liệu và Quy trình Xử lý

1. Mục tiêu xử lý dữ liệu

Dự án tập trung thử nghiệm trên một lĩnh vực pháp luật nhỏ để kiểm chứng hiệu quả của hệ thống hỏi đáp. Lĩnh vực được chọn là **Luật Hôn nhân và Gia đình** tại Việt Nam, với tập dữ liệu thử nghiệm bao gồm:

- Các câu hỏi trắc nghiệm thuộc các mức độ dễ, trung bình, và khó liên quan đến Luật Hôn nhân và Gia đình.
- Các văn bản pháp luật và nội dung giải đáp từ nguồn đáng tin cậy.

2. Phạm vi dữ liệu

Lĩnh vực: Luật Hôn nhân và Gia đình.
Nguồn dữ liệu:

Websites uy tín: Các trang web pháp luật chuyên ngành như **Thư Viện Pháp Luật**, **LuatVietnam**, với nội dung bao gồm:

- Điều luật cụ thể trong các văn bản pháp luật.
- Các bài viết phân tích liên quan.
- Hệ thống câu hỏi và câu trả lời từ chuyên gia pháp lý.

Văn bản luật chính thống:

- Luật Hôn nhân và Gia đình năm 2014 cùng các văn bản hướng dẫn liên quan.
- Các văn bản pháp luật được tổng hợp từ **Thư Viện Pháp Luật** và các nguồn tin chính thức khác.

Sách và tài liệu: Giáo trình, tài liệu pháp luật chuyên ngành về Luật Hôn nhân và Gia đình.

3. Công cụ thu thập dữ liệu

Web Crawler:
Dữ liệu được thu thập bằng công cụ **Tavily Search**, với các bước thực hiện:

Thu thập nội dung từ web: Lấy toàn bộ nội dung văn bản, tóm tắt các ý chính và thu thập các siêu dữ liệu liên quan như: Ngày đăng, Tác giả, Tên nguồn trang web.

Kết quả thu thập: Tổng cộng **1.321 websites** đã được xử lý.

4. Quy trình xử lý dữ liệu thô

Xử lý dữ liệu thô:

Loại bỏ thông tin không liên quan:

- Loại bỏ các đoạn nội dung không thuộc lĩnh vực Luật Hôn nhân và Gia đình.
- Lọc bỏ các thông tin thừa, ví dụ: quảng cáo, nội dung lặp lại, hoặc không liên quan đến pháp luật.

Chuẩn hóa dữ liệu:

- Định dạng lại nội dung văn bản để phù hợp với hệ thống hỏi đáp.
- Gắn nhãn và phân loại dữ liệu theo các tiêu chí như: câu hỏi, điều luật, phân tích, giải đáp từ chuyên gia.

Xây dựng tập dữ liệu sạch:

- Tóm tắt và tổ chức nội dung để làm đầu vào cho hệ thống RAG và LLM.

Tạo tập data đánh giá

Các câu hỏi thực tế:

Để đánh giá hiệu quả hệ thống hỏi đáp pháp luật trong lĩnh vực **Luật Hôn nhân và Gia đình**, dự án đặt mục tiêu thu thập và tổ chức bộ câu hỏi bao gồm:

1. **Bài kiểm tra trắc nghiệm**:
    - Lựa chọn một bài kiểm tra cấp đại học gồm 50 câu hỏi trắc nghiệm, có đáp án và phần giải thích chi tiết (nếu có).
2. **Bài kiểm tra tự luận**:
    - Thu thập một bài kiểm tra tự luận với các câu hỏi mở gồm 30 câu, kèm theo đáp án chi tiết để làm tiêu chuẩn đối sánh (gold standard).
3. **Câu hỏi thực tế**:
    - Tổng hợp các câu hỏi pháp lý từ các trường hợp thực tế hoặc câu hỏi gửi đến chuyên gia pháp luật. Gồm 20 câu.

Điểm số được chấm dựa trên các tiêu chí sau:

- 0 điểm: Hệ thống không tìm nạp (retrieval) được thông tin và trả về phản hồi dạng "Tôi không được cung cấp thông tin...".
- 1-2 điểm: Hệ thống tìm nạp được một số thông tin nhưng hoàn toàn không liên quan hoặc sai so với câu trả lời đúng.
- 3-5 điểm: Tìm nạp được một số thông tin chính xác, hữu ích nhưng chưa đầy đủ và vẫn chứa thông tin không chính xác.
- 6-7 điểm: Tìm nạp được hơn 50% thông tin đúng, thông tin không chính xác đã giảm nhưng chưa loại bỏ hoàn toàn.

- 8-9 điểm: Tìm nạp được 70%-90% thông tin đúng, phần thông tin không liên quan hoặc sai sót không ảnh hưởng lớn đến câu trả lời.
- 10 điểm: Tìm nạp đầy đủ tất cả nội dung có trong câu trả lời đúng, không thừa, không thiếu, dù diễn đạt có thể khác.

Sinh câu hỏi để đánh giá:

Mục tiêu: Thiết kế bộ câu hỏi yêu cầu hệ thống phải hiểu và xử lý toàn bộ dữ liệu thay vì chỉ trả lời các câu hỏi chi tiết, cục bộ.

Các bước thực hiện:

- Bước 1: Mô tả ngắn gọn tập dữ liệu
  - Bao gồm các thông tin như: tập dữ liệu là gì, chứa nội dung gì, mục tiêu sử dụng của tập dữ liệu.
- Bước 2: Xác định người dùng tiềm năng và nhiệm vụ
  - Với mỗi tập dữ liệu, hệ thống LLM được yêu cầu xác định 5 người dùng tiềm năng và mỗi người thực hiện 5 nhiệm vụ cụ thể.
- Bước 3: Tạo câu hỏi
  - Dựa trên từng cặp (người dùng, nhiệm vụ), hệ thống LLM tạo ra 5 câu hỏi cần khả năng tổng hợp thông tin toàn diện từ tập dữ liệu.
  - Kết quả cuối cùng: Mỗi tập dữ liệu có 125 câu hỏi kiểm tra.

Tiêu chí đánh giá

Hệ thống được đánh giá dựa trên 4 tiêu chí chính:

- Comprehensiveness (Toàn diện): Mức độ chi tiết và đầy đủ của câu trả lời trong việc giải quyết mọi khía cạnh của câu hỏi.
- Diversity (Đa dạng): Sự phong phú và đa chiều trong cách trình bày, cung cấp các góc nhìn khác nhau về câu hỏi.

- Empowerment (Khả năng hỗ trợ quyết định): Mức độ hữu ích của câu trả lời trong việc giúp người đọc hiểu vấn đề và đưa ra các nhận định chính xác.
- Directness (Tính trực tiếp): Mức độ rõ ràng và cụ thể trong việc trả lời câu hỏi

Công cụ sử dụng:

- Một hệ thống LLM đóng vai trò "giám khảo" được cung cấp câu hỏi, tiêu chí đánh giá và hai câu trả lời từ hệ thống RAG.

Phương pháp thực hiện:

- Với mỗi câu hỏi và tiêu chí, hệ thống LLM so sánh hai câu trả lời và:
    1. Chọn câu trả lời tốt hơn theo từng tiêu chí.
    2. Trường hợp hai câu trả lời gần như tương đồng → Kết quả được ghi nhận là hòa (tie).
- Để giảm thiểu sự ngẫu nhiên trong đánh giá, mỗi cặp câu trả lời được so sánh 5 lần.

Giảm thiểu sai lệch:

- Để đảm bảo tính công bằng, thứ tự trình bày hai câu trả lời trong prompt được thay đổi luân phiên.
- Tính toán tỷ lệ chiến thắng (win rate) cho từng tiêu chí, từ đó tổng hợp kết quả cuối cùng.

Kết Quả

Bộ database gồm có:

- 2772 chunks
- 2738108 tokens

- 1949688 từ
- 14289 thực thể (nút)
- 15774 cạnh → Mật độ đồ thị = 2E/V(V-1) = 0,015%
- 75 câu hỏi từ các bài thi
- 125 câu hỏi được sinh ra.

Kết quả đánh giá, baseline là naïve RAG:

- 125 câu hỏi được sinh ra: tỉ lệ thắng là ~60%
- 75 câu hỏi từ các bài thi: điểm trung bình của cả hai phương pháp đều là khoảng 8 điểm

V. Kết luận

Kết quả của các câu hỏi rộng, khó cho thấy GraphRAG trả lời tốt hơn, toàn diện hơn.

Kết quả của các câu hỏi local thì các hai phương pháp đều có khả năng trả lời như nhau.

Tuy nhiên ở một số tiêu chí Naïve RAG vẫn có kết quả tốt hơn so với GraphRAG là vì **i)** yếu tố lớn nhất ở đây là dữ liệu chưa được nhiều, chưa bao quát **ii)** trong quá trình xử lý dữ liệu còn thiếu xót, chưa lọc được hết dữ liệu dư thừa.

Tóm lại, phương pháp đã cho thấy khả năng tốt, hiệu quả khí có bộ dự liệu lớn và thời gian truy xuất, trả lời nhanh, hợp lý.

VI. Tài liệu tham khảo

LightRAG: https://arxiv.org/pdf/2410.05779

GraphRAG Survey: https://arxiv.org/pdf/2408.08921

LangGraph: https://langchain-ai.github.io/langgraph/

ReAct: https://arxiv.org/pdf/2210.03629

Reflexion: https://arxiv.org/pdf/2303.11366